Corpws Cenedlaethol Cymraeg Cyfoes

National Corpus of Contemporary Welsh

Contents

Greetings from the PI

Welcome to the 14th issue of the CorCenCC newsletter. It has been yet another busy couple of months at CorCenCC HQ and across all the partnering institutions. In this edition, we aim to bring you up to date with recent progress of the project including some hot off-the-press news about an exciting new project partnership and details of a successful funding bid, developments which both promise to strengthen our work even further. We will also update you on some of the events that we have participated in recently, including Swansea University's programme in the UK wide 'Being Human' festival and our inaugural mini CorCenCC away day. The latter of these brought together some members of the team based in South Wales for an interactive team-building and bonding event held on a beautifully autumnal day near the shores of Swansea Bay. Building on September's edition we also bring you updates from the final two work packages on the project, WP2 and WP5, and another short report about the experiences of one CorCenCC researcher 'On the Road' (this time featuring Jennifer Needs). Our final 'feature' is this month's meet the team profile which includes…well …me! As it is the final edition of 2017 we all want to wish you a very Merry Christmas and a Happy New Year. The newsletter will be back in January - we hope to see you again then!

Happy reading! Dr Dawn Knight

BBC Cymru/Wales: new CorCenCC partners

The CorCenCC team are proud to announce that an agreement has been signed between the project and BBC Cymru/Wales, making them partners on the project. This partnership will enable us to use content from BBC Cymru Wales within CorCenCC (including BBC Cymru Wales' TV and radio programmes and also online articles). We want to say a big thank you to BBC Cymru/Wales for their collaboration – we look forward to working with them on the project!

www.bbc.com/wales

BBC cymru wales

News and events

The CorCenCC team are pleased to announce that we have been awarded funding from the Welsh Government's Grant Cymraeg 2050 scheme for work on a project entitled WordNet Cymraeg.

The aim of the project is to automatically construct a WordNet for Welsh, a lexical database in which words are grouped into sets of synonyms (synsets), which are then organised into a network of lexico-semantic relationships. WordNets are widely used in natural language processing (NLP) to support understanding of meaning expressed in written and spoken language. As such, WordNet is vital for language technology applications such as question answering, information retrieval and machine translation.

These technologies are vital for development of user-friendly interfaces of smartphone and smart home apps, which will drive the use of Welsh-medium digital technology for Cymraeg 2050.

By linking the WordNet Cymraeg project to the CorCenCC project, we will re-use its sustainability and engagement plans to increase the visibility and ensure the long-term future of the WordNet Cymraeg.

Public engagement activities include:

Llywodraeth Cymru
Welsh Government

- A social media campaign will be carried out to advertise the project and encourage users to test functionalities.
- Regional road shows/workshops will be held at schools, libraries and community centres/ Mentrau Iaith to raise potential users' awareness of the WordNet and to provide basic training of its utilities.

WordNet Cymru will be led by Professor Irena Spasic, working with Dr Dawn Knight and Dr Steven Neale

New team member!

We want to take this opportunity to introduce you to the newest member of the CorCenCC team, Lowri Williams. Lowri will be working as a part-time Administrative Assistant for the duration of the CorCenCC project and will be based at Cardiff University. *"Hello, I'm Lowri and I joined the project in October as Admin Assistant. My background is in financial services, having worked for almost 10 years at Legal & General on their IT projects. While there I pursued my secret lifelong obsession: language and linguistics. I completed my degree with the Open University this summer and I'm now enrolled on an MA at Cardiff. It was during a meeting with Dawn Knight a couple of years ago that I first heard about the CorCenCC project and signed up to do some transcription. I consider myself amongst a silent majority of 'lapsed' South Welsh speakers: learned it at school but haven't had much chance to use it since. As a Welsh speaker, I think it's great that the project includes speakers of all levels and walks of life. As a linguist, I'd be particularly interested to see how Welsh has adapted in response to modernisation and the spread of English. I'm a self-confessed 'geek', and when I'm not at the Uni in some capacity I'm mad about games and puzzles. This August a couple of my friends and I formed a quiz team for the BBC2 show Only Connect. We didn't win (robbed!) but it was a great day out and it's something I can now cross off my (embarrassingly nerdy) bucket list. I'm genuinely excited by what the project can bring to Welsh speakers and learners alike, and I look forward to helping in whatever small way I can. Diolch!"*

On Friday 17 November, Jenny Needs and Steve Morris went to the Ty'r Gwrhyd 'Canolfan Gymraeg' in Pontardawe to hold the only Welsh medium event at this year's 'Being Human' Festival. This is the UK's only national festival of the humanities and the only hub the festival has in Wales is in Swansea. The festival is led by the School of Advanced Study, University of London in partnership with the British Academy and the Arts and Humanities Research Council.

The name of the CorCenCC Welsh medium session was "Rho dy Gymraeg i ni / We want your Welsh!" and it was a fantastic opportunity to collect hours of spoken data through experimenting with the 'Gogglebox' television programme model and asking participants to give a live reaction to short films.

Swansea University
Prifysgol Abertawe

Being HUMAN 17 – 25 NOVEMBER 2017

CorCenCC runs the only Welsh medium event in the 2017 'Being Human' Festival

17th to 25th November, Swansea University

It was also an opportunity to engage with the public in the Swansea Valley and show them the app (as Jenny is doing in the picture). There was a good – and lively – response to the films from the 'sofa critics' and this is definitely a way of collecting data which we will look to use again in the future.

CorCenCC Away Day

6th November 2017, Swansea University

"it gave the team some much needed time to catch-up face-to-face, to take stock and positively reflect on the progress we have made so far"

On the 13th of November, the WP1 team (the PI, Co-Investigators and RAs) met up in Swansea for an away day. This meeting was multifunctional; it gave the team some much needed time to catch-up face-to-face, to take stock and positively reflect on the progress we have made so far, and to prioritise and plan the remaining months of the project. The meeting was productive and a good way to introduce new members of the team. To break the ice, our communication skills were tested by completing interactive communication tasks using images - in which we completed in record time. But the main focus was on how to stream line handling Big Data, and identifying the challenges of data collecting and possible ways of limiting them. Overall, the outcomes were positive, and we have already started implementing changes to data collection methods, such as the use of a web scraper to automate the extraction of e-language texts.

Lowri Williams, Cardiff University

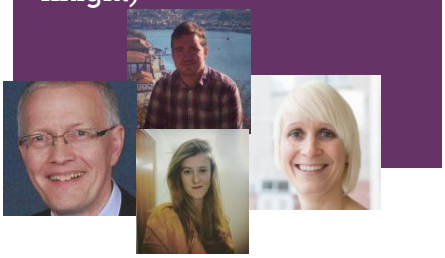
Work Package (WP) 2 and 4 updates

Our work on WP2 has really taken shape over recent months. Our bespoke part-of-speech tagging software, *CyTag*, is now complete, robust, and working really well. We've also managed to put together a gold standard evaluation set of 600+ sentences manually annotated with part-of-speech tags, which has allowed us to properly evaluate *CyTag*. We're very happy with the results, which are excellent for Welsh and are comparable with the type of accuracy expected from part-of-speech taggers in other languages.

One of the things we're most proud of with *CyTag* is that it leverages open-source materials to help in the process of deciding on parts-of-speech. It works primarily by using the information found in Kevin Donnelly's *Eurfa* – the largest open- source, freely available dictionary for Welsh – to produce a list of possible tags for each word in a Welsh text. This is supported by specific lists of place names, first names and surnames extracted from Wikipedia data. Once a list of possible words has been produced, we're then able to use a set of bespoke rules to prune the possible tags for each word – based on the tags or features of its neighbouring words – until we arrive at the correct one. For example, in the sentence 'mae Cymru yn wlad Geltaidd' ('Wales is a Celtic country') we can assume that 'yn' is a particle (linking 'Cymru' and 'wlad') because we know that 'wlad' is a soft mutation of 'gwlad' ('country'), and we have a rule to select the particle tag for 'yn' if the following word is a soft-mutated noun. See how important it is to correctly deal with mutations?

WP2:

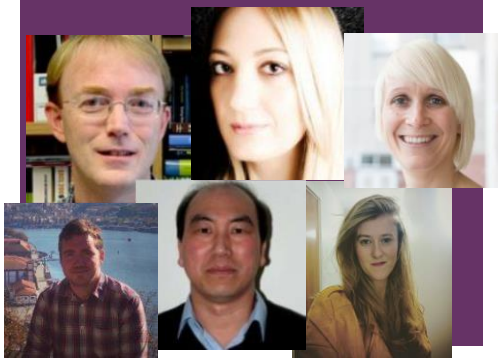
Develop the part-of-speech tag-set/tagger (lead: Dawn Knight)



We plan to build on our progress over the next couple of months with some additional features and extensions to *CyTag*. We'll be looking at whether machine learning techniques can help us to lemmatise (find the uninflected root form of) Welsh words when they're not found in Eurfa, and we'll also begin to look at how to tag multi-word expressions (as opposed to solely single-word forms). Keep an eye out too for our *CyTag* website with online demo and more information, which will be available shortly.

WP5:

Construct infrastructure to host CorCenCC and build the corpus (lead: Irena Spasić)



WP5 deals primarily with the technical infrastructure for the project, and we've made good progress recently in two key aspects of this. Our most interesting developments have been in the early stages of setting up our corpus query tools themselves – we've been working with a small dataset to ensure that we're going to be able to run all of the queries we want on it once it's been tagged using *CyTag* and our proposed semantic tagger, and are now shifting our attention to the design of the interface itself to make sure it looks right for CorCenCC's users. We're hopeful of having a working demo available in the new year, so that we can then start populating with data as the corpus grows.

Secondly – and with the more functional running of the project in mind – we've been streamlining the processes used to record and manage our collected data. We've

provided web interfaces for researchers and transcribers to move files around in storage and share access with each other, and are currently working on further interfaces for categorising that data according to the project's sampling frame. We've also implemented a streamlined database design for recording information about data collected, allowing us to keep a closer eye on what's been collected so far.

Work on the CorCenCC Crowdsourcing App also falls under the umbrella of WP5, and we've been working recently on getting our Android version ready to complement the already available iOS version. We're also working on overhauling a web version of the app, which gives contributors another way of sending us data if they don't have access to the bespoke apps. Concrete news on the availability of all versions of the App will be available early in the new year.

+ CorCenCC on the road

At the start of the month, I spent a week 'in the field', meeting Welsh speakers and recording them going about their daily lives through the medium of Welsh. Sometimes people think CorCenCC is looking for particular examples of Welsh, and they refer us to the 'best' speakers of the local dialect. But what we are really after is examples of how the language is used by the general public – whilst shopping, whilst socialising, over a cuppa or over lunch. I hope that the diary below (an especially busy day during my week away) will show some of the things we are trying to collect across the country – and perhaps make you think which examples of Welsh you could contribute yourself!

07:30 – I arrive at the first recording location of the day – a leisure centre, where I will be collecting examples of conversations between staff and customers at the reception. I'm there very early, but a lot of the customers have beaten me! However, there's a decent flow of customers going in and out, and I'm lucky that this is quite a Welsh-speaking area, so most of the conversations take place through the medium of Welsh.

09:00 – Breakfast time! I'm in another village now and choosing a café where the staff speak Welsh, in the hope that they will be willing for me to record conversations in the café. But they're short-staffed today, so there isn't even a chance for me to mention the CorCenCC project. A quiet breakfast for me today, then. Better luck next time!

10:00 – I'm ready for my next appointment – in a library, this time. My week is a mixture of organized appointments (with companies, organisations, and individuals), and trying to seize opportunities that arise between appointments. At the library, I record conversations between staff and customers and between members of staff as they go about their work. Usually, I avoid being the recordings myself – including too much of my own language would skew the data, as we're trying to sample enough speakers to reflect Wales' diversity. But I make an exception today – a couple of the library customers are very interesting and want to share their stories with me, so I have to be in the conversations! I try my best to limit my responses to "ooh" and "mhm", instead of skewing the data with too much talking!

11:30 – I'm 'on the road' again, for a recording appointment in the next town. I'm intending to collect examples of the language of clients and staff, but the reception is very quiet today. I'll have to come back sometime when the place is busier. But I recorded a couple of conversations between staff, so that's something!

13:30 – I've asked for permission to record at another café over lunch. I arrive late because I fell and twisted my ankle! I record people ordering food whilst I have a quick cuppa, but it's time for me to move on to the next appointment! But I'll be back at the café again in the new year to get more data.

14:00 – A group of older people from the area meet every week to catch up over a cuppa. This time, they have come to the local school, where the pupils provide sandwiches, cakes and tea. Everyone's enjoying themselves and I record sections of the lively conversations that are flowing like the tea! A very lovely occasion to be part of – thank you!

"Everyone's enjoying themselves and I record sections of the lively conversations that are flowing like the tea! A very lovely occasion to be part of – thank you!"

Jennifer Needs

16:00 – I go back to the B&B to transfer the recordings to the laptop and to sort paperwork, before having supper and going out to the last recording session of the day.

19:30 – And now for something completely different! Tonight I'm recording a public lecture – a complete change from the language I've been collecting during the day. We are trying to collect a bit of everything, so here's an example of the use of slightly more formal Welsh. The speaker is funny and very interesting. But remember – you don't have to be funny OR interesting to give us your Welsh!! Contributions of any kind – formal or informal, interesting or mundane – will help create a picture of what Welsh really is today.

Jennifer Needs, Swansea University

+ Meet the team: Dr Dawn Knight, Principal Investigator

So, who is Dr Dawn Knight, PI on the CorCenCC project - the one behind the light-bulb hair and oft-tooth-filled-grin? Well now is the time to find out, in the final 'Meet the Team' for 2017.

I probably should start by answering a (not so) simple question: where am I from? Well, my dad was in the REME (Royal Electrical and Mechanical Engineers) when I was a child, so I spent most of my time (until the age of 11) living near army barracks in Paderborn, Germany. This makes me an original 'Brit Brat' - or so I am told. The first picture sums up those early years pretty well. Some people say I haven't changed much since then…

We then moved across the channel to Bovey Tracey, Devon where I spent my more formative years running through fields on my grandparents' farm, climbing up tors on Dartmoor and scrumping apples. During those early years I was convinced that I would do something maths-related when I

was 'older', but, instead, I developed a love for Literature during my sixth form years and decided to enrol on a degree in English Studies at the University of Nottingham in 2000. As the first in my family to go to University, I had no idea what to expect – and certainly no idea that I would never escape those ivory towers again.

It was during my BA that I was exposed to the field of applied linguistics and I guess I was immediately hooked. In 2003 I chose to continue studying and completed an MA, then progressed on to the PhD. This was followed by a post as an RA (Research Assistant) on a couple of large-scale ESRC (Economic and Social Research Council) funded projects on constructing multimodal and heterogeneous corpora. I was very privileged to work closely with Svenja Adolphs and Ronald Carter during my latter years in Nottingham. They really helped me to develop my understanding of the value and potential for using corpus-based methods to answer a range of different 'real-world' questions about language as it is used in different contexts, cultures and across different times and modes of communication (including gesture – my PhD work focused on backchanneling head nods, for example). They also helped to develop my skills in presenting and communicating ideas effectively, which proved equally invaluable for me. Believe it or not, I started my PhD as a very shy, retiring individual who only shouted on a football pitch on Sunday afternoons (when playing for Nottingham Forest Ladies). After a few years of careful nurturing and mentoring, I soon 'flowered' into a more confident human, and a more able (perhaps institutionalised) academic. They have a lot to answer for, I guess!

I flew my Nottingham nest up to Newcastle to take up my first post as a lecturer in Applied Linguistics in 2011.

It was a nippy but nice time up North. It was in the Toon where the (perhaps more relevant) CorCenCC story begins. I am sure many of you have heard this by now. If not, it went a little like this: one too many espressos in a Toon greasy spoon + Tess Fitzpatrick + a madcap Twister Welsh-language learning game idea = the foundations for building a contemporary corpus of Welsh. With Steve Morris on board for a coffee fuelled second meeting, held in Bristol Temple Meads station, a fuller, more exciting, behemoth of a research plan was formulated. I hadn't met Steve before that meeting, but after those few hours it felt a little like we'd known each other forever. We had fun and to this day we continue to have a fabulous, quirky partnership and a great friendship. After 3 years of planning, drafting and redrafting, the proposal was submitted, and I moved south to Cardiff University. We had a good feeling about the bid, but I don't think I will ever forget the day when we were told we had been successful – it suddenly all became real.

We were really heartened by the fact that, during the planning phases, all potential collaborators, partners and stakeholders near-immediately understood the importance and potential of what we had envisioned with CorCenCC.

The associated impact of the work seemed to really strike a chord with everyone we involved. In fact, I have to admit that I feel so lucky to be part of the CorCenCC (dream) team: we really do have some inspirational, intelligent, hard-working and visionary people involved. The unfaltering enthusiasm of individuals was particularly encouraging for us, given the amount of time, energy and belief Tess, Steve and I had already put into the plans for the project. Belief is perhaps the most important part of the whole equation and is something that I feel, to this day, drives the entire team through the research. While it is very hard work at times, as projects of this scale and magnitude often are, it is this belief that will hopefully make CorCenCC a success both when it draws to a close in 2019, but well into the future too.

We want CorCenCC to be a living, breathing beast, and while it will take a lot of work to keep the beast alive, we will all do our very best to make sure we succeed. January marks the 3rd year anniversary of my arrival in Wales and in my time here I have learned a lot about the country, Welsh culture and language. It has been a fun-packed, thrilling, fast-paced and, at times, intense journey so far…long may it continue.

+ 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