Corpws Cenedlaethol Cymraeg Cyfoes

National Corpus of Contemporary Welsh

Greetings from the PI

Welcome to the 19th edition of the CorCenCC project newsletter. September has arrived and the start of the new academic session has begun. We are now officially into the final year of the project – where has the time gone?! Within the next 12 months the corpus will be launched and all plans will finally come into fruition. What will you search the corpus for first?

In this issue of the newsletter we will bring you news of some summer events and conferences; the latest developments of the crowdsourcing app; some insights into working on the project from our summer placement students and last but not least, will introduce you to members of the extended CorCenCC family.

Happy Reading – Dr Dawn Knight

Contents

P1: News and events
P3: CUROP reflections
P4: My CorCenCC PhD
P5: Meet the team
P6: Contact us

+ News and events

App now on Android!

Remember our crowdsourcing app? To complement the readily available iOS version, we've recently been working on creating more ways for Welsh speakers to include their data to the corpus. Now you can also contribute data to us via an overhauled web version, or on Android platforms!

One of our goals on CorCenCC is to include data directly from the community in the corpus, in order to ensure that contemporary Welsh is well-represented in the 10 million words. The app gives Welsh speakers the opportunity to record conversations between themselves and others across a range of contexts so that they can be included in the final corpus, and to upload metadata that helps us to categorise such contributions. Crowdsourced data is a relatively new direction, and so we see this as an exciting way to gather contemporary Welsh in addition to more traditional language data collection methods (which you'll be familiar with if you've already seen CorCenCC team members out and about).

Of course, the key ingredient in all of this is your data. Whether it's chatting with your friends in the pub or the café, your family at the dinner table, or anything else you can think of that you'd like to share, your Welsh is important to us – and we want to see as much of it as possible reflected in CorCenCC and available for the community to see and to learn from. So please do consider taking part as a contributor, or helping us spread the word about the availability of the app. CorCenCC will be a resource for all, so let's make sure that it reflects all of our Welsh. Our crowdsourcing app is one way to do that. If you're interested in contributing via the app, you can register and contribute online at http://app.corcencc.org, or download our iOS or Android versions using the QR codes here. Please also feel free to email us with any app-based questions at tech@corcencc.org.

Recent events: National Eisteddfod of Wales, 3-11 August 2018

This year's National Eisteddfod has been described in the media as 'experimental', as it was located in the centre of Cardiff Bay rather than in a muddy field. We spoke to a great number of people during the week who thought that the experiment had paid off: the lack of boundaries meant that anybody could stumble across the Eisteddfod so the festival felt open to everyone, not just the usual Eisteddfod-goers. From the CorCenCC team's perspective too, it was great to see so many people around – and that so many of you were willing to contribute a conversation to the corpus, over lunch, coffee, a pint or a gin! A very big 'Diolch' to you all!

We also had success recording at a number of stalls at the Eisteddfod. Customer-staff language can sometimes be difficult for us to collect, because we can't predict whether Welsh-speaking customers will visit a shop on the day that we are there to record. But at the Eisteddfod of course, there were customers everywhere, and almost every one of them a Welsh speaker/learner. (Remember that the corpus will include the language of learners and of other speakers who are less confident with their Welsh – there's no expectation for contributors to speak perfectly. **We want your Welsh!**) We asked Sioned Johnson-Dowdeswell, a SPIN student from Swansea University, for her thoughts and experiences of working at the Eisteddfod with the CorCenCC team: 'Working with the CorCenCC team during the National Eisteddfod this year was a very valuable experience for me. I found the task of explaining the project to members of the public rather challenging at times, because the work is initially quite difficult to understand. But with practice and experience, it became easier to gain people's attention and interest. I met a large number of interesting people at the Eisteddfod, and their enthusiasm for the project and the Welsh language was heartening. I am not a naturally confident person, therefore approaching and communicating with total strangers was quite an achievement for me! I feel that I have benefitted personally by undertaking this work as I have gained in confidence whilst collecting relevant and useful data'.



Recent events: BAAL annual conference 2018 @ York St John

Members of the CorCenCC Management Team, Dawn Knight, Tess Fitzpatrick and Steve Morris travelled to York St John to attend the annual British Association for Applied Linguistics Conference (6th – 8th September). 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. The team presented two papers based on the CorCenCC work. Responding to the theme of the conference, 'Taking Risks in Applied Linguistics', the first of the papers provided a candid reflection of some of the challenges and rewards associated with planning, securing, managing and disseminating a large-scale research project. The presentation offered some practical guidance and top-tips emerging from the lessons we've learned for others

who are seeking, or encouraged, to lead and/or contribute to large-scale research projects. The second gave participants and opportunity to see some CorCenCC's digital tools in operation. This included a demonstration of the semantic and POS taggers; a pilot version of the pedagogic toolkit and some worked example of the basic corpus query tools in operation. Both presentations were well attended with some great feedback, and interesting questions, from those in attendance of the talks. As members of the Executive Committee for BAAL, the conference is always a very busy time for the CMT, but it always proves to be an intellectually stimulating and rewarding conference. The conference dinner and entertainment were fabulous, and we even had a chance to have a quick look around the cultural and historical hotspots of the beautiful city of York. Steve and Dawn even bumped into some morris dancers (Steve was in his element!). The next BAAL conference coincides with the end of the project, so we will be back in attendance, at Manchester Metropolitan University, to dot the i's and reflect back on how the entire project has gone.

CUROP reflections by Alys Greene, Cardiff University

So, I have now reached the end of my 8 week placement working on the CorCenCC project, and how the time has flown by! I have managed to do much more than I would have imagined, and the work has been very worthwhile. From my first couple of weeks familiarising myself with the world of corpus linguistics and understanding why corpora are created to appreciating what my contribution on WP3 would achieve and why the semantic tagger is needed and how it works in practice, it has been a valuable experience.

My work focused on the semantic tagging of the data, where data was given semantic tags to explain the context in which it was used by the tagger, and I would then manually check that the correct tags had been applied. I soon got to work on the tagging, using the Welsh USAS Semantic Tagset to manually tag the files and make changes where they were needed to give the data the right meaning. Some files proved to be more challenging than others, but the more familiar I became with the Tagset the easier the tagging became.

I was then lucky enough to have a short break from the office and got to experience another side of the project; going out data collecting on the maes at the Eisteddfod. It was very different to what I was used to doing in the office, but was a good chance to get talking to people about the project and we managed to collect social and transactional spoken data.

It was then time to get back to the office and finish tagging the files I had left. I have now finished those tasks, and having enjoyed my time on the project, look forward to seeing where it goes from here. Being a part of the CorCenCC team has been a great experience, I've been able to work independently on my own tasks and also with a great group of co-workers on other aspects of the project, and it has also given me the opportunity to work through the medium of Welsh, something I hope to continue to do after finishing my studies. As for the future, I hope to stay involved and help with the project in any way that I can, whether it be through contributing more data or transcription. Watch this space!

My CorCenCC PhD - by Bethan Tovey, Swansea University

Bethan Tovey recently started her CorCenCC-affiliated PhD at Swansea University. We asked Bethan to give us some details on what she plans to study for her PhD, the importance of this work and how it 'adds' to the general aims of the CorCenCC project:

Although reliable figures are hard to find, it seems likely that more of the world's population is multilingual than not. Yet in Britain, we have a tradition of aggressive monoglottism that manifests itself in a linguistic chauvinism directed scattershot at anyone whose English is not "native" enough. Its other targets, of course, are the pre-English languages of Britain, like Welsh.

As Welsh-speakers, we're all too familiar with the same old jokes about Welsh, trotted out daily on social media, and often enough in the national media, as well. One of the common tropes is to mock Welsh speakers for introducing English words into their Welsh. According to one Twitter user, "50% of the welsh language is taking English words and changing the vowels to 'i's an [sic] 'w's".

Welsh-speakers are also prone to complaining about this . Another Tweet complains "It annoys me when people use "Welsh-English" when we have perfectly good words in y Gymraeg e.e. trowsus vs llodrau, miwsig vs cerddoriaeth, jam vs cyffaith... I know language is a living thing but we have such a beautiful one that it doesn't need to be bastardised".

Switching between languages is a phenomenon known by names including "code switching" and "code mixing". Research across a range of multilingual communities suggests that this is a common linguistic feature amongst people who have more than one language in common. It can fulfil any number of important strategies, including creating interpersonal closeness or distance, or emphasising key information.

For some speakers, however, code mixing feels like a failure, especially when they switch from a minority or heritage language to a dominant language such as English. As yet another Twitter user puts it, "You're worried that your grammar's poor, that you drop too many English words in, and you don't 'sound like a Welsh speaker.'"

My PhD research will consider Welsh-English code mixing in its cultural context, looking at whether and how personal attitudes towards the phenomenon affect individuals' linguistic behaviour. Unlike much of the existing research, I'll be studying mixing in both directions, recording data from people speaking Welsh and speaking English, in order to see whether there is asymmetry in the way code mixing manifests itself in these different contexts.

Code mixing will naturally be a feature of the CorCenCC data. Because Welsh speakers are almost all also English speakers, when we speak together we know that we will still be understood if we mix Welsh with English. This is not the case for perhaps the majority of other language pairings: an English-German bilingual cannot know, when speaking with an English-speaking stranger, whether code mixing with German will be appropriate or intelligible. As a result of the status of Welsh as a bilingual language, as it were, Welsh-speakers' speech, informal writing, and electronic communication tends to contain a range of code mixing behaviour, from single-word borrowings, to switching between phrases, sentences, and larger units.

Amongst other things, this has implications for Welsh-medium education. For example, if we can discover whether certain topics of discussion are more likely to trigger frequent code mixing, we might advise

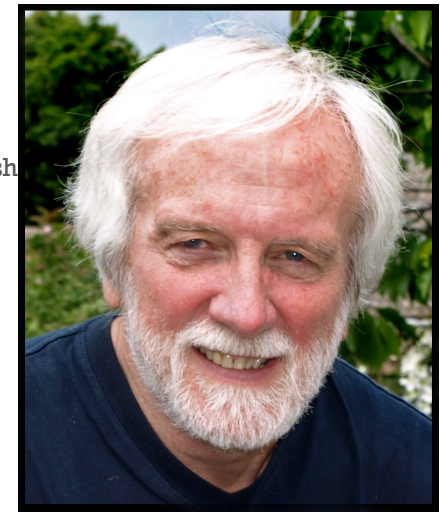
examiners to avoid those topics, since they are unlikely to elicit the most traditionally "correct" performances in Welsh from students. English-medium educationalists are coming to understand the importance of taking students' dialects of English into account, and helping them acquire Standard English as a professional tool rather than insisting they replace their variety of English with the standard. A similar movement might be of benefit in Welsh education, to recognise that some students, for whom code mixing is a strong feature of their dominant (or only) variety of Welsh, may need help to acquire an academic variety of Welsh that avoids code mixing alongside their usual dialect.

The purpose of my research, however, is not to encourage people to avoid code mixing. Rather, one of its aims is to raise awareness of how natural a behaviour code mixing is for bilinguals, and to use that awareness to shape appropriate curricular and teaching aims for both schoolchildren and adult learners. I also hope to encourage those who believe that their Welsh is "not good enough" because of code mixing. Finally, I hope to develop our understanding of how Welsh affects a bilingual speaker's use of English, encouraging a model of Welsh-English bilingualism as a two-way exchange rather than an adulteration of one language by the other.

+ Meet the team: Professor Michael McCarthy, CorCenCC project consultant

I was born in Cardiff and learnt Welsh at school in the way that it was taught in those days (the late 1950s, early 1960s), that is, more like the way we learnt Latin than the way were learning French or Spanish. We learnt word lists, we learnt the tables of mutations by heart (I can still recite most of them today!), we translated sentences and short texts and learnt to conjugate verbs. Having said that, it wasn't all dull, dreary rote learning; we also learnt beautiful poems which we had the chance to recite and compete with in school Eisteddfodau. I found my metier in that activity and often scored highly. And on one occasion, I was given the great and unforgettable opportunity to sing in Welsh in the youth choir at the National Eisteddfod when it was held in Cardiff. The only problem was we never learnt to speak the language or have conversations in it, so it was all a rather abstract subject, which put a lot of pupils off, though not me.

For one reason and another, I ended up studying Spanish at university and my travels as an English teacher took me to Spain for nine months and to Sweden for five years, so Swedish and Spanish dominated and Welsh got kicked out, so to speak, from the warehouse of my mind. In short, it got so rusty it almost vanished forever.



CorCenCC has been a great stimulus to me to renew my knowledge of Welsh and I keep making resolutions to go on a refresher course, always seeming to be thwarted by other pressures and responsibilities. I can understand quite a lot of the discussion in Welsh at CorCenCC meetings but can't contribute in Welsh; more's the pity. But as a corpus linguist, I know just how important the project is and I will serve it in whatever way I can, albeit through the medium of English.

+ 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, Alex Lovell and Jonathan Morris; **RAs** - Steven Neale, Jennifer Needs, Mair Rees, Scott Piao and Lowri Williams; the **PhD students** - Vigneshwaran Muralidaran and Bethan Tovey; **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