CorCenCC Newsletter

Issue 7: October 2016

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Greetings from the PI

Welcome to the seventh edition of the CorCenCC newsletter. Now that the conference season is over, we are beginning to shift our attention to perhaps the most important task for the project: data collection. In the coming weeks and months, members of the team will be working hard to recruit participants and begin amassing the 10 million words of spoken, written and digital (e-language) Welsh that will make up the CorCenCC corpus [if you would like to help us with this do let us know!]. In this newsletter, we introduce you to the CorCenCC crowdsourcing application that will help us to achieve this, by allowing users to record their own spoken language. The app will be launched officially in the next few weeks.

In this newsletter we also provide some more general updates from the 5 key project work packages, and we reveal some exciting news about a recent agreement made between CorCenCC and S4C. Finally, in our 'meet the team' section we introduce you to Tess Fitzpatrick, Professor of Applied Linguistics in the Centre for Language and Communication Research and CorCenCC Co-Investigator (CI).

Happy reading, Dr Dawn Knight (Cardiff University)

News

The CorCenCC team are proud to announce that an agreement has been signed between the project and S4C, making them partners on the project. This partnership will enable us to use some data from the channel within CorCenCC (from TV programmes to meeting recordings), and help with public engagement activities and events. We want to say a big thank you to S4C for their collaboration – we look forward to working with them on the project!

S4C

www.s4c.cymru/cy/

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)

Over the last few months we have been working hard to complete our sampling framework. This has been important work as it will inform our data collection strategy and ultimately the corpus content. We are now on the cusp of collecting Welsh language data, and the entire team is much enthused about this. If you want to contribute your written, spoken or e-language data please do get in touch!

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

Our gold-standard manually annotated corpus is well under construction, with annotators working hard to assign tags from our bespoke part-of-speech tag-set to a small selection of Welsh text. We'll use this corpus in the coming months to evaluate the processing tools that we are also developing for this work package.

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

In WP3, work is underway to compile a large-scale Welsh semantic lexicon for the Welsh semantic tagger. So far, we have collected over 329,000 Welsh words from a variety of corpus resources, which was found to cover 97.63% of the words in a test corpus.

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

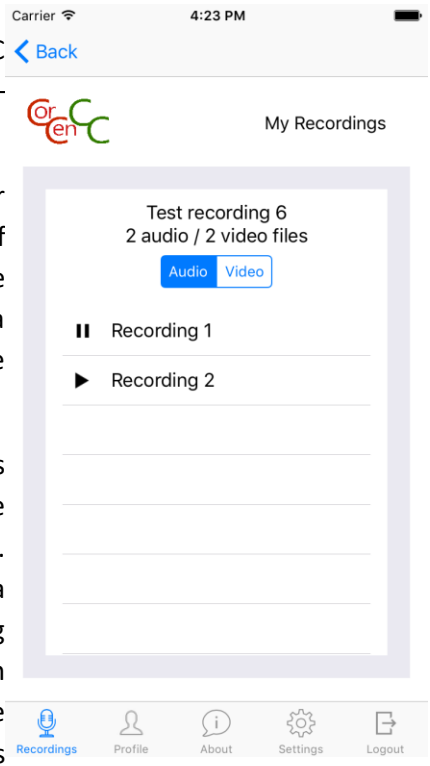
The online version of our questionnaire was launched in September, and we have succeeded in reaching Welsh teachers, trainers, lecturers and tutors from a wide variety of contexts – from those who teach at primary schools (Welsh-medium and English-medium) to those who teach adults. Many thanks to everyone who has already responded – your input will help us to tailor our pedagogical toolkit according to your needs and the needs of your students. If you haven't had a chance to complete the questionnaire yet, it's not too late! Please follow this link http://arolwg.corcencc.cymru/holiadur/holiadur-i-addysgwyr/ to share your knowledge and experience with us.

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

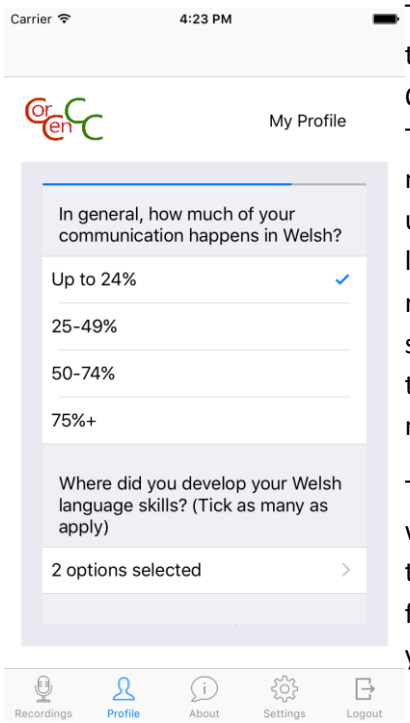
Our CorCenCC Crowdsourcing Application is now ready and about to be released for technical testing. We plan to launch the application in the coming weeks, ready to let it loose in the field to collect your Welsh language data!

CorCenCC Crowdsourcing Application

This month we finished work on the first beta version of the CorCenCC Crowdsourcing Application – or 'app' as they are more commonly known – which is now being tested in advance of its upcoming first public release.



The app – currently for iOS but with an Android version planned for the near future – allows Welsh speakers to record their conversations across a range of contexts so that they can be included in the final corpus. It's going to be one of the first apps of its kind for this purpose, with crowdsourced corpus data being a relatively new direction to complement more traditional language data collection methods.



The app makes this process easy and means that Welsh speakers can engage with the CorCenCC project at their own convenience. This makes contributing to the corpus a much more personal experience, giving users ownership and control of their own language data rather than having it be mediated in any way. By making the process seamless and by putting speakers themselves in charge of their contributions to the corpus, speakers can be sure that the recordings they share will be the most natural and accurate representation possible of their everyday Welsh.

The app is free to use and a first version will be officially launched in the coming weeks following the testing period. You can keep track of the app's launch via the usual CorCenCC channels (Facebook, Twitter and our newsletter) and please feel free to contact us for more information if you are interested in contributing your Welsh language data using the CorCenCC Crowdsourcing Application.

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 Tess Fitzpatrick, who is a Co-Investigator on the CorCenCC project, based at Cardiff University.

Profile: Tess Fitzpatrick

In 2011 I gave a talk at Newcastle University about how language learners make links between the words they know and the words they learn. Dawn Knight hosted the event, and in a café afterwards, she and I plotted a project to create a kinaesthetic vocabulary learning device for children, and I suggested we pilot it in Welsh primary schools. (That project, working title "twister/dance mats", is still on the cards.) On the train home I received a text from Dawn: 'we'll need to use the Welsh corpus – can you send me a link?'. I consulted Steve Morris, who I worked with at Swansea, and we sent Dawn various links, but it was immediately apparent that no comprehensive, accessible, searchable corpus of Welsh existed. A faint vision of CorCenCC started to shimmer on the Tyneside and Swansea Bay horizons.

"the beginning of the CorCenCC story..."

Having worked, tangentially, with both Dawn and Steve, I was aware not only of their expertise (in corpus linguistics and Welsh language/pedagogy, respectively), but also of the energy, commitment and ready humour with which they approach their work – and I think all three of us would agree, five years later, that the last of these has been a crucial part of the mix. It seemed to me that if a national Welsh corpus was to be created, these were the people to make it happen. A half day meeting between the three of us in Bristol Temple Meads station café confirmed the skill, knowledge (and humour) "fit" of the core team, and the vision of CorCenCC was coming into focus; after two more years of meetings in various places (a Southampton canteen, a Bristol hotel foyer, my kitchen in Swansea), we had a full project team and a completed grant application – the vision was clear.

"where the CorCenCC bid was cooked up....."

CorCenCC makes sense to me in ways that relate to my own experience as a language teacher, language learner, applied linguist – and resident of Wales. Between 1989 and 2003 when I worked as an English language teacher and teacher trainer, multimedia teaching and self-access materials (many of which were corpus-informed) were the standard tools of my trade. I realised that not all language teaching was so well-resourced, but this understanding came from calls for spare coursebooks for developing countries, and, in the early 1990s, from the training courses I led for teachers in eastern Europe who had to meet a new, urgent demand for English classes. In the late 1990s I began studying part-time for a PhD in Applied Linguistics, and suddenly I had a whole new perspective on language teaching and learning. One impact of this was a significant addition to my interpretation of "well-resourced", which can be summarised as "supported by tools and materials that are in turn informed by evidence-based research", and this defines, for me, the contribution that CorCenCC can make to Welsh language teaching. I am in the pedagogic work package team, and knowing how my own experience as a practitioner

underpins my research in second language acquisition, I am excited about creating a "tool-kit" that is truly user-informed.

I moved to Wales from England in my early 20s; I grew up in rural Kent, and moved to London at 18. My awareness of the value of living in a bilingual community grew very gradually, and without much initial attention – I lived here for 20 years before going to a Welsh class. Increasingly, I have become fascinated by the intricacy of the connections between language and community, the subtle influence of the bilingual landscape, and the sense of access to new domains that I have experienced from my first, stuttering attempt at Welsh (which was a slow, careful "paned o de" to Mark Stonelake's "beth ti'n moyn" in a café – I wanted coffee but panicked). I feel privileged to be living in a bilingual part of the world, and to be working with the (multilingual) CorCenCC team.

"Exploring the (linguistic?) landscape of Wales"

Tess Fitzpatrick, FitzpatrickT@cardiff.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