CorCenCC Newsletter

Issue 11: May 2017

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Greetings from the PI

Welcome to the eleventh CorCenCC newsletter! In this edition, we update you on how the project work has progressed, across each individual work package (WP), as well as updating you on some of our plans for the summer months. It is that time in the year when the CorCenCC team begin to step up their attendance at academic conferences (the 'conference season') and public events, so there are many exciting times ahead! This month we will also bring you some 'reflections from the inside' with the researchers from Work Package 1 (WP1) providing some thoughts on what they have found, to date, to be the most interesting, fun and challenging experiences of collecting Welsh language data for CorCenCC. This is complemented by an interview with one of the people who has kindly donated data to the corpus. Hopefully, this will help you to understand how easy it is to 'get involved' with the project and will inspire you to contribute to CorCenCC too! As with previous editions, we will introduce you to yet another member of the extended CorCenCC family. This month is the turn of Steven Neale, an RA based at Cardiff University. Oh, and if you have any ideas about future items/content for this newsletter, or would like to get involved in putting it together, do get in touch!

Happy reading, Dr Dawn Knight (Cardiff University)

Updates from CorCenCC Work Packages (WPs 1-5)

Work on the project is distributed across 6 coordinated work packages (WPs), 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)

During recent months, we have started to get to grips with the task we were all looking forward to so much, that is recording our fellow Welsh speakers as they use their Welsh. So far, we've been collecting language throughout Wales, from Carmarthenshire in the South, to Anglesey in the North, and lots of places in-between as well! We intend to re-visit these places over the next few months, and plan to visit some new places as well of course. Keep an eye on our Facebook page and our Twitter account to find out when we will be visiting your area. Talking about spoken language, we have also appointed a number of transcribers so that we can store the recorded spoken language in the form of text in the corpus.

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

We've almost completed work on our bespoke part-of-speech tagger, which has been under development for a while now. We've been checking the output of the tagger in recent weeks to get an idea of its accuracy, and of course to find out where we can make improvements to it in future iterations. Keep an eye out for the release of a web-based demo version shortly!

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

With regards to the development of the Welsh semantic tagger, the Welsh semantic lexicons have been further expanded using more lexical resources, particularly the TermCymru database which is a Welsh-English translation term bank compiled and published by the Translation Service of the Welsh Government. The recent focus for WP3 has been on extracting Welsh multiword expressions (MWE) and identifying their correct semantic categories via the bilingual

term bank. Meanwhile, a web demo site (http://phlox.lancs.ac.uk/ucrel/semtagger/welsh) and a desktop user interface (GUI) have been developed for the current version of the Welsh semantic tagger for people to try and test it. In the coming weeks, the team will work on the further development of the Welsh semantic tagger by testing and incorporating more efficient word sense disambiguation methods and exploring more corpus and lexical resources.

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

In our last update (issue 7), we mentioned our online questionnaire which will help us to tailor our pedagogical toolkit according to the needs of teachers and students. In addition to the questionnaire, we have been meeting teachers and tutors face-to-face, to raise awareness of the corpus and the pedagogical toolkit, and to continue to collect views that will help shape the toolkit's development. It has been wonderful to see practitioners' enthusiasm for the corpus and what it could offer for teachers and learners, both inside the classroom and out. It's a real privilege to be able to work on something with such potential to benefit the future of Welsh learning!

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

Work is underway on the development of our front-facing corpus query tools, which people will eventually use to search for words, structures and information in CorCenCC. As data starts coming in from other work packages, we'll be able to populate these front-facing tools and people will soon be able to start making use of the resource and exploring contemporary usage of Welsh! We've also been working on the development of an Android version of the CorCenCC Crowdsourcing Application, which is coming along nicely.

Future plans, conferences and events

Over the coming weeks, the WP1 team will continue to visit counties across Wales to collect data. This work includes recording at a number of events, from internal meetings, public lectures, choir practices and coffee mornings, to public events and festivals such as Sesiwn Fawr Dolgellau and the Royal Welsh Show's Spring Festival – not to mention the Urdd Eisteddfod and National Eisteddfod!

But data collection is not the only work going on this summer. CorCenCC team members will be presenting the project's work at the Corpus Linguistics 2017 international conference as well! Over the five days of the conference, we will be giving four presentations and presenting a poster too, in order to raise awareness of the project's different work packages amongst other researchers. Details of the papers follow:

- Piao, S., Rayson, P., Watkins, G., Knight, D. and Donnelly, K. (2017). Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language. Paper to be presented at the *Corpus Linguistics Conference 2017*, July 2017, University of Birmingham.
- Rees, M., Watkins, G., Needs, J., Morris, S., and Knight, D. (2017). Creating a Bespoke Corpus Sampling Frame for a Minoritised Language: CorCenCC, the National Corpus of Contemporary Welsh. Paper to be presented at the *Corpus Linguistics Conference 2017*, July 2017, University of Birmingham.
- Neale, S., Spasić, I., Needs, J., Watkins, G., Morris, S., Fitzpatrick, T., Marshall, L., and Knight, D. (2017). The CorCenCC Crowdsourcing App: A Bespoke Tool for the User-Driven Creation of the National Corpus of Contemporary Welsh. Paper to be presented at the *Corpus Linguistics Conference 2017*, July 2017, University of Birmingham.
- Needs, J., Knight, D., Morris, S., Fitzpatrick, T., Thomas, E.M. and Neale, S. (2017). "How will you make sure the material is suitable for children?": User-informed design of Welsh corpus-based learning/teaching tools. Paper to be presented at the *Corpus Linguistics Conference 2017*, July 2017, University of Birmingham.

➢ Knight, D., Fitzpatrick, T., Morris, S., Evas, J., Rayson, P., Spasić, I., Stonelake, M., Thomas, E.M., Neale, S., Needs, J., Piao, S., Rees, M., Watkins, G., Anthony, L., Cobb, T.M., Deuchar, M., Donnelly, K., McCarthy, M. and Scannell, K. (2017). Creating CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes - The National Corpus of Contemporary Welsh). Poster to be presented as part of the CMLC-BigNLP2017 National Corpora Poster Track at the *Corpus Linguistics Conference 2017*, July 2017, University of Birmingham.

Reflections 'from the inside'

To give you a better idea of what it is like to be a CorCenCC team member 'in the field' (i.e. helping to collect data), we've asked two of the researchers on WP1 to reflect on the following questions:

- What have been the highlights of the data collection process for you so far?
- What sort of things have been cool/interesting and what have been the key challenges?
- What are your immediate plans for data collection?

Jennifer: *"Recording data for CorCenCC has given me the opportunity to attend a whole host of events I wouldn't otherwise have experienced: I had a very warm welcome at a poetry class in West Wales, and again at public speaking competitions in the North and the South. The experiences were very different, but all very interesting – I really enjoyed myself! And very soon I'll be visiting Mid Wales and attending the Royal Welsh Show's Spring Festival! These are all quite unique examples of how Welsh is used in Wales today, but one of the best things about the data collection work is the opportunity to hear Welsh being spoken in everyday situations – in shops and cafés, for example. I met a couple of ladies who had come to a café to catch up over lunch, and when I explained to them about the CorCenCC project and asked for permission to record them, they were really enthusiastic about the idea of the corpus and totally happy to contribute to it. They had a very natural conversation, with the recorder just running in the background. Perfect! (If you'd like to contribute a similar conversation, get in touch! Remember that we change any names that come up, so you don't have to avoid gossip!) I was also invited to record a couple of families in their homes – excellent opportunities again to record very natural language. It can be difficult to find such opportunities, though, unless the researcher already knows a member of the family. That's one of the reasons why the CorCenCC app is such a great idea – it gives people the opportunity to record their own conversations, without one of the CorCenCC researchers having to be there (affecting the ambience!). There are over half a million Welsh speakers in Wales, and we can't invite ourselves over to every single one's homes! But it's really easy for people to download the app and contribute that way – we'd love to see lots of people having a go at using the app, since collecting natural everyday language is a really important part of our work. (Have you tried the app yet? It's available via the Apple Store, and an Android version is on its way!)"*

Gareth: *"The highlight of the data collection so far? Well, actually, every recording session is a highlight. Hearing the Welsh language being used, and more than that, flourishing in our cities, our towns, our villages, our homes, our cafes and our (hic) pubs is a wonderful experience, a privilege. Convincing a handful of people to let me record them has been challenging. Some believe that their language is not good enough (not true – we want to collect EVERYONE's language). Others are shy. But usually even the most shy of people forget that the recorder is in front of them and chat away contentedly. I'm looking forward to meeting yet more Welsh speakers over the coming weeks and months as I visit Neath Port Talbot, Rhondda Cynyon Taf, the Vale of Glamorgan and Bridgend."*

Siân contributes to CorCenCC - Be like Siân!

Who are you?

I'm Siân Griffiths, Chief Officer of Blaenau Gwent, Torafen and Mynwy's Menter Iaith. The Menter is based is St James Hall, down in Pontypool.

What does the Menter Iaith do?

*We are the only Menter Iaith which is responsible for three counties, so we have plenty of work to do! We concentrate mainly on developing Welsh in the community for the benefit of children, young people, families, learners and anyone who is keen to use or develop their Welsh lang*uage skills

How did you become aware of CorCenCC?

Well, Wales is a small country, and as it happens, I had met one of the CorCenCC research assistants previously, in the Welsh for Adults field. When she started in her new role, she got in touch and came over to explain the aim of the project. Since then, we have been working closely together. I've been emailing her reports and so on, so that the text can be added to the corpus. She's also been attending events arranged by the Menter, so that she can record those present speaking. Ultimately, these conversations will be transcribed and also added to the corpus

Why did you decide to help out?

In my opinion, the project is an important one which is going to support the Welsh language in the future. I can see that it has great potential to be the basis for research into the development of the language by comparing the use of words, dialects and so on. At the end of the day, that will benefit us all. Also, the corpus is going to be a very useful resource for tutors and people learning the language. And of course, we have a high proportion of people who are learning Welsh as a second language in this area.

Will the corpus have any particular advantages for people in you area*?*

As I said at the start, I'm the chief officer of a Menter Iaith which is responsible for three counties, two of which contain disadvantages areas. Obviously, the fact that this resource is going to be available online, free for everyone to use is a great thing for us here.

If you'd like to have a chat about contributing to the corpus, sent us an email at corcencc@caerdydd.ac.uk and Jenny, Gareth or Mair (our research assistants) will get back to you. Look forward to hearing from you.

Menter Iaith
Blaenau Gwent•Torfaen•Mynwy

Menter Iaith BGTM: Contact details
Rhif ffôn: 01495 755861
Ebost: gwybodaeth@menterbgtm.org
Gwefan: www.menterbgtm.cymru
Twitter: @MenterBGTM1
Facebook: https://www.facebook.com/menterbgtm/

Canolfan Iago Sant - *St James' Hall*
Heol Hanbury - *Hanbury Road*
Pont-y-pŵl - *Pontypool*
Torfaen, NP4 6JT
01495 755861

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 Dr Steven Neale, who is an RA on the CorCenCC project, based at Cardiff University.

Profile: Steve Neale

I do like these 'Meet the Team' sections. All these people that I work with day-to-day, or see pretty regularly at the various project meetings, and there's always something interesting to discover about their background that I didn't know before. Writing my own is pretty similar – taking a moment to think about how I ended up where I am has left me sitting here thinking 'did I really do that? Surely that all happened to a different person?'. I guess you could say I've taken a fairly unconventional route to get to this point (and thus, to CorCenCC), but I'll try to keep it short(ish) and sweet…

I grew up in Leicestershire, with no intention to do anything else other than work in TV and film. My first degree was in Media Production, and after filming a very low-budget Match of the Day-style football 'piece' I landed my first job, as a video producer filming interviews and editing match highlights at Nottingham Forest FC. I spent a season at close-quarters with the team, attending more games than I ever thought possible with my camera in-hand and trying to keep my love for their local rivals (Derby County) under wraps, before leaving to take what I thought was my dream job working in post-production at a film and broadcast company in central London. But after realising that it wasn't really for me, I went back to De Montfort University in Leicester, got my masters in Creative Technologies, and – inspired by my favourite lecturer from those Media Production days – made the decision to pursue an academic career.

First stop – PhD. I'm still not quite sure how I ended up moving to Tasmania, Australia to do it, but there you go. I loved academia – problem solving, managing my research tasks, even writing the papers – but I was still yet to find my 'calling', as such. I'd decided to work on Computing because I knew my way around a computer from my media and creative technology days, and I felt like when I found a theme that really captured me I could push things in that direction. Little did I know that I'd been discovering which way I was going to push things in my own time – on a whim, I'd started learning Spanish just before starting the PhD, somehow ended up surrounded by 'hispanolhablantes' both in Tasmania and during a 5-month research exchange in Germany, and started not only making real strides with it but becoming hugely passionate about it as well.

It wasn't long before I realised that *languages + computer science* ≈ *natural language processing* (NLP). I made the risky decision to try and land a post-doc in NLP – full of motivation but totally lacking in experience - after wrapping up my PhD, and am eternally grateful to António Branco, who took a chance on me and gave me my first post-doc at the University of Lisbon in Portugal in a field I was completely new to. It worked out well (hopefully he'd say the same), and over 18 great months in the Portuguese capital I'd found my niche, got papers out and experience under my belt, and was doing something I really loved. It even encouraged me to fully embrace my 'techier' tendencies – I found I quite liked working on the command line and writing shell scripts if the output was language-related.

And so, CorCenCC. After so much time abroad in so many great locations, I don't think anywhere but Wales – with the vibrancy and interest there is here for the language – could have tempted me back to the UK. The chance to work on NLP problems in Welsh is exciting enough from a professional standpoint, and as somebody interested in language learning the opportunity to be exposed to Welsh to this level on a day-to-day basis is also really exciting.

But I think one of the biggest attractions to CorCenCC for me is its position as a resource for learners and teachers (among many others), and having benefited immensely from learning languages myself, it's great to work on something that I'll no doubt be querying myself when I forget which preposition to use when buying Welsh cakes or a pint of Brains in the coming years.

Steven Neale, NealeS2@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 new 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