CorCenCC Newsletter

Issue 8: November 2016

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Greetings from the PI

So here we are – CorCenCC's eighth newsletter, signalling the end of the ninth month of work on the project. Time sure does fly when you are having fun! Last month we informed you of the near-completion of the crowdsourcing app – thanks to numerous volunteers this has now been tested extensively and the final touches/improvements on this are currently being made. We expect the app to be available in the Apple app store in the very near future so do keep your eyes peeled (don't worry, an Android version will be on its way sometime in 2017 too)! The completion of the app dovetails with perhaps the most time-consuming aspect of the project: data collection. As you know, we aim to collect at least 10 million words of Welsh language-in-use by the end of the project. To give you a clearer sense of the types/varieties/genres of the data to be included, and the nature of the participants that we hope to source this from, the next three newsletters will provide an overview of the 'sampling frame' i.e. data collection plans for the e-language, spoken and written data respectively. Hopefully this will help to fill you in on the 'bigger picture' of the project and will entice you to get involved and contribute your Welsh! Also in this month's edition, we will bring you the latest news from the project and we will, again, introduce you to a member of the team (Laurence Anthony).

We also wanted to let you know that this will be the final edition of the newsletter for the year. From now on we will be producing a newsletter on a two-monthly basis (so expect the next one sometime in January). The newsletters will continue to provide you with up-to-date news, progress and other information on the project – just in a fun-packed 'bumper' format. So now seems like an appropriate time for us to say Season's Greetings to one and all – we [team CorCenCC] hope you all have a fabulous Christmas and we look forward to catching up with you in the New Year!

Happy reading, Dr Dawn Knight (Cardiff University)

News

In February 2017 Tess Fitzpatrick, co-investigator on the CorCenCC project, will be moving from one of our partner universities to another! She will be leaving Cardiff to take up the post of Head of English Language and Applied Linguistics at Swansea University, where she will join CorCenCC team members Jenny Needs, Mair Rees, Mark Stonelake and Steve Morris. Tess says *"Although I will miss working on a daily basis with the excellent colleagues in Language and Communication at Cardiff, the fact that the project team spans both universities will make the transition much easier. I'm looking forward to the challenges and opportunities that my new role will bring, and to working with new colleagues and contexts. The project will be key in establishing and enhancing common ground between Welsh Language and Applied Linguistics activities at Swansea, as we continue to build and strengthen links between the project's academic and community partners".*

We want your Welsh: A focus on e-language

As you know, a key aim for the CorCenCC project is to create a corpus which is balanced and represents all forms of Welsh as it is 'actually' used on a day-to-day basis. This means that it will include language from a range of different types, discussing different topics, and from a variety of different contributors from all walks of life. Two

million of the ten million words that we aim to collect will be sourced from 'e-language' resources. These are essentially communicative modes that are digitally born – e.g. the language typed into your phone, computer or any other electronic device. As the world is becoming increasingly digital it is important that the corpus captures and represents how language is being used in this way.

We focused on collecting e-language data back in 2013 when we initially piloted the idea of CorCenCC, so we have had some practice in recruiting e-contributors and collecting their data. Back then we managed to build a mini corpus with over 500k words from 72 different people. For CorCenCC 'proper' we are planning to extend this by collecting around 600k words each from blogs and websites, and 400k words each from emails and SMS messages from hundreds of Welsh users. We were initially hoping to sample tweets as well, as we understand that Twitter is well-used in the Welsh language context. However, Twitter's Rules of the Road limit the number of tweets which can be shared with third parties. As all of CorCenCC's outputs will be freely available to members of the public, any such limit is not compatible with the CorCenCC ethos, so unfortunately it won't be possible to include tweets in our corpus.

As an initial guideline for sampling the data, we have decided to target the collection of the public-facing e-language (i.e. websites and blogs) according to 6 broad thematic groupings. These groupings are based on a similar classification system used in another e-language corpus: CANELC (the Cambridge and Nottingham E-Language Corpus – Dawn was the research assistant who worked on collecting data for this corpus back in 2009/2010), although some of the categories have been modified in order to make the system appropriate to the Welsh language context. The groupings range from more formal topics in the public sphere (e.g. news, media and current affairs) to more informal and personal/individual topics (e.g. parenting and family life).

In the case of the more private types of e-language (i.e. emails and SMS messages), we obviously won't know anything about the data until we receive it, so instead of sampling these forms of data based on themes, we will target contributors based on the purpose of the communication, that is whether they are communicating about personal or business matters. This is broadly in line with the sampling frame suggested for spoken data (more on this in the next issue!) in that business messages are more likely to be Transactional or Professional in nature, while personal messages are more likely to be categorised as Socialising or Private language. It is likely that business short electronic text messages will be fewer, while business email messages will be more plentiful, so our sampling frame's word targets have been designed with this in mind. Below is the sampling frame for e-language data. Again, this is just a guideline for what we want to collect – an 'ideal'. In reality the distribution of data in CorCenCC is likely to be quite different to this, but this acts as a useful starting point/foundation for us to build on. Please take a look and, if you would like to contribute any of these types of language data to the corpus, please get in touch!

Thema/Pwnc / *Theme/Topic*	**%**	**Geiriau/*words***
Blog	**30%**	**600,000**
A: Newyddion, Y Cyfryngau a Materion Cyfoes / *News, Media and Current Affairs*, Gwleidyddiaeth / *Politics*, Busnes a Chyllid / *Business and Finance*, Y Tywydd a'r Amgylchedd / *Weather and the Environment*, Siopa Ar-lein / *Online Shopping*	5%	100,000
B: Crefydd / *Religion*, Yr Iaith / *Language*, Diwylliant, Llenyddiaeth a'r Celfyddydau / *Culture, Literature and the Arts*, Addysgu, Academia ac Addysg / *Teaching, Academia and Education*	5%	100,000
C: Technoleg, Cyfrifiaduron a Chwarae Gemau Cyfrifiadurol / *Technology, Computers and Gaming*, Ffasiwn a Harddwch / *Fashion and Beauty*, Hobïau a Difyrrwch / *Hobbies and Pastimes*, Teithio / *Travel*, Coginio / *Cookery*	5%	100,000
D: Cerddoriaeth / *Music*, Chwaraeon / *Sport*, Perfformiadau byw a Digwyddiadau / *Gigs and Events*	5%	100,000
E: Hynt a Helynt Pobl Enwog / *Celebrity news and gossip*, Teledu a Ffilm / *TV and Film*, Hiwmor / *Humour*	5%	100,000
F: Bod yn Rhiant a Bywyd Teuluol / *Parenting and Family Life*, Iechyd a Lles / *Health and Wellbeing*, Bywyd Personol a Phob Dydd / *Personal and Daily Life*	5%	100,000

Gwefan / *Website*	30%	600,000
A: Newyddion, Y Cyfryngau a Materion Cyfoes / *News, Media and Current Affairs*, Gwleidyddiaeth / *Politics*, Busnes a Chyllid / *Business and Finance*, Y Tywydd a'r Amgylchedd / *Weather and the Environment*, Siopa Ar-lein / *Online Shopping*	5%	100,000
B: Crefydd / *Religion*, Yr Iaith / *Language*, Diwylliant, Llenyddiaeth a'r Celfyddydau / *Culture, Literature and the Arts*, Addysgu, Academia ac Addysg / *Teaching, Academia and Education*	5%	100,000
C: Technoleg, Cyfrifiaduron a Chwarae Gemau Cyfrifiadurol / *Technology, Computers and Gaming*, Ffasiwn a Harddwch / *Fashion and Beauty*, Hobïau a Difyrrwch / *Hobbies and Pastimes*, Teithio / *Travel*, Coginio / *Cookery*	5%	100,000
D: Cerddoriaeth / *Music*, Chwaraeon / *Sport*, Perfformiadau byw a Digwyddiadau / *Gigs and Events*	5%	100,000
E: Hynt a Helynt Pobl Enwog / *Celebrity news and gossip*, Teledu a Ffilm / *TV and Film*, Hiwmor / *Humour*	5%	100,000
F: Bod yn Rhiant a Bywyd Teuluol / *Parenting and Family Life*, Iechyd a Lles / *Health and Wellbeing*, Bywyd Personol a Phob Dydd / *Personal and Daily Life*	5%	100,000

Ebost / *Email*		20%	400,000
Proffesiynol *Professional*	e.e. ebost i gadarnhau amser cyfarfod *e.g. an email to confirm a meeting*	13.4%	268,000
Personol *Personal*	e.e. ebost sy'n rhannu newyddion da *e.g. an email to share good news*	6.6%	132,000

Negeseuon Testun Electronig Byr / *Short Electronic Text Messages*		20%	400,000
Proffesiynol *Professional*	e.e. Neges sydd wedi ei hanfon gan ysgol sy'n darparu gwybodaeth ynghylch noswaith rhieni / *e.g. a message sent by a school providing details of a parents' evening*	6.6%	132,000
Personol *Personal*	e.e. neges ynghylch cwrdd â ffrind am goffi *e.g. a message regarding meeting a friend for coffee*	13.4%	268,000
		100%	**2,000,000**

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 Laurence Anthony, who is a Project Advisory Group member on the CorCenCC project, based at Waseda University.

Profile: Laurence Anthony

When I got invited to write this introduction to my work and connection to the CorCenCC project, my first instinct was to go back to all the previous "Meet the team" introductions to get an idea about the style, length, and topics that would be considered. I soon found that the others in the team used on average 268.8 different word forms in their introductions, and that the introductions were on average 538.4 words in length. I also found out that some of the key words in the introductions were "Welsh", "language", "resources", "developing", "my" and "I". So, now I could start writing! Yes… for better or worse, that's me… a corpus linguist!

So, let me explain my background and why I became involved in the CorCenCC project. I grew up in Huddersfield to a mother from Cardiff and father from London. When I was growing

up, all I wanted to be was an astrophysicist. I loved the idea so much that I read all the books by physicists like Stephen Hawking and Richard Feynman (even naming my son Richard!), learned how to build electronic circuits and program computers, and even went to university to study the subject. But, while studying at the University of Manchester Institute of Science and Technology (UMIST), my career path took a dramatic turn when I went on a trip to Japan and met many scientists and engineers there who were struggling to communicate in the foreign language of English. I decided then that after graduating, I would move to Japan and focus my career on teaching scientific communication and developing educational and research resources to help people learn and analyze language.

Twenty-five years on, I'm doing the same kind of work. But, one thing has clearly changed over that time. The longer I have lived outside of the UK, the more I have appreciated the value, importance, and insights gained from knowing more than one language. So, I have focused much of my time on developing tools that support research, teaching, and learning of multiple languages. In fact, one of my most popular software programs is a freeware corpus analysis tool called *AntConc* that is used by people in over 140 countries around the world. In view of this, perhaps it's easy to understand why I jumped at the chance to serve as a consultant on the CorCenCC project, where I can support the CorCenCC team in the creation of Welsh language research analysis tools and learning resources.

From my perspective as a corpus linguist, the CorCenCC project is ground-breaking in many ways. For example, in order to collect authentic Welsh language data, the project will be using cutting-edge crowd-sourcing methods that will allow us to see how the language is used in real, daily settings. Also, because the scale of the project is so large, the data will have to be stored in a unique way so that people can access the huge number of results quickly and easily. The project is also ground-breaking in the way that it carefully links the collection of language data with the development of practical teaching and learning tools and materials. So, it's really a honor to be part of the project and help more people understand, learn, and teach the Welsh language.

Now… this introduction has reached 559 words, so I should probably stop here!

Laurence Anthony

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