CorCenCC Newsletter

Issue 9: January 2017

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Greetings from the PI

Welcome to the ninth edition of the CorCenCC newsletter. 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, over the course of three newsletters we will provide an overview of the 'sampling frame' i.e. data collection plans for the e-language (issue 8), spoken language (this issue) and written language (next issue). 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 (Steve Morris).

Happy reading, Dr Dawn Knight (Cardiff University)

News

We are pleased to welcome a new member to the CorCenCC team, Vigneshwaran Muralidaran! Vignesh has just begun a PhD at Cardiff University, the work on which will dovetail with some of the activities on the CorCenCC project. Vignesh is being supervised by Dr Dawn Knight and Professor Irena Spasić. We invited Vignesh to tell us all a little bit about himself: *I graduated with a Masters degree in Computational Linguistics from International Institute of Information Technology, Hyderabad (IIIT-H) in 2016. During these years I worked at the institute as a Research Assistant in Language Technology Research Centre (LTRC) and as a teaching assistant for Computational Linguistics course. My research was on formulating a Construction Grammar based theoretical framework to parse Dravidian languages and development of a full parser for Tamil. I earned my Bachelors degree in Information Technology from Amrita Vishwa Vidyapeetham, Coimbatore in 2010 and subsequently worked as a programmer analyst in industry for three years before joining IIIT-H. Currently I am interested in exploring the implications of treating language as a functional, usage-based system and how it can be exploited in Language Technology*.

We want your Welsh: A focus on spoken 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. Four million of the ten million words that we aim to collect will be sourced from spoken resources.

CorCenCC's spoken language data will be collected from 7 main contexts of language use in Wales, ranging from those associated with more formal language such as institutional contexts and television/radio to those associated with less formal language such as private and social contexts. Our 'sampling frame' for spoken language – the framework that will guide us as we choose what to record – has been drawn up very carefully, taking into consideration the numbers of people who regularly produce certain types of language (e.g. more people use Welsh privately and socially than professionally) as well as audience size of certain types of language (e.g. television and radio language). Our aim is to reflect the diversity and richness of the Welsh language, not just by collecting a token amount of each type/genre, but by collecting an *appropriate* amount of each, so as to create a 'snapshot' of our language as it is used today.

Our 7 spoken language contexts are based on data collection models used for the British National Corpus and CANCODE (the Cambridge and Nottingham Corpus of Discourse in English). The genres and sub-genres within each context have been tailored specifically for Welsh, and therefore include language used at Eisteddfodau, language from Welsh-medium education and from classes of Welsh as a 2nd language, and a relatively large target for language from children's TV programmes. As well as trying to ensure an appropriate balance of the different genres, CorCenCC aims to appropriately represent Welsh speakers themselves too. This means that our research assistants will be travelling all across Wales – to rural areas as well as towns and cities – and recording speakers of all ages and with all sorts of language backgrounds. Whether you have spoken Welsh all your life, whether you learned the language through Welsh or English-medium education, or whether you've learned Welsh more recently or indeed are learning it now, **we want your Welsh!** All Welsh speakers contribute to the mosaic that makes up the Welsh-speaking community, and CorCenCC would love to be able to reflect this.

Below is the sampling frame for spoken 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! **There are two main ways you can contribute spoken language data: (1) contact us to find out when the CorCenCC research assistants will be recording in your area, or (2) record yourself using our brand new app!**

Genre	Sub-genre	%	Words
Public/Institutional		**10%**	**400,000**
Political speeches and public lectures	e.g. electoral broadcasts, speeches by MPs/AMs/councillors, speeches by the Welsh Language Society/Friends of the Earth/etc., public lectures	3.75%	150,000
Parliamentary language	e.g. parliamentary proceedings	1.25%	50,000
Formal language at church/chapel	e.g. sermons/homilies, prayers	3.75%	150,000
Formal language from the National Eisteddfod	e.g. Music/Dance/Recitation/Drama adjudications, presentations at university stands/in the Lolfa Lên/Maes D/etc.	1.25%	50,000
Media		**15%**	**600,000**
TV programmes (BBC/S4C)	Children	3.375%	135,000
	Entertainment	1.275%	51,000
	Factual	1.25%	50,000
	Drama	0.9%	36,000
	Sport	0.9%	36,000
	Music	0.275%	11,000
	News	0.275%	11,000
Radio programmes (BBC Radio Cymru)	Entertainment	1.375%	55,000
	Factual	1.375%	55,000
	News	1.1%	44,000
	Music	0.825%	33,000
	Religion & Ethics	0.275%	11,000
	Sport	0.275%	11,000
	Comedy	0.275%	11,000
Commercial/community radio programmes		1.25%	50,000

Transactional		10%	400,000
Transacting goods, information and services (face-to-face or phone calls)	Public sector, e.g. councils, Welsh Language Commissioner's office, post offices, HMRC, libraries	3.5%	140,000
	Retail, e.g. shops, auctions	1.5%	60,000
	Food, e.g. restaurants, cafes	1.5%	60,000
	Tourism & travel, e.g. tourist information, accommodation, national rail enquiries	1.5%	60,000
	Leisure, e.g. media, clubs, entertainment, beauty	1.5%	60,000
	Finance, law & health, e.g. accountants, solicitors, clinics, pharmacies, hospitals	0.5%	20,000

Professional		10%	400,000
Workplace Welsh	Staff meetings	2.5%	100,000
	Interviews	2.5%	100,000
	Desk-to-desk conversation	2.5%	100,000
	Business-to-business interaction (face-to-face or phone calls)	1.25%	50,000
	Internal phone calls	1.25%	50,000

Pedagogical		10%	400,000
Primary classroom interaction	Welsh-medium classroom interaction from a variety of age groups	3.5%	140,000
Secondary classroom interaction	Welsh-medium Science and Maths lessons	1%	40,000
	Welsh-medium Art, Drama & Music lessons	0.375%	15,000
	Welsh-medium Geography, History & Religious Studies lessons	0.75%	30,000
	Welsh-medium Welsh lessons	0.875%	35,000
Secondary and Further Education (FE) classroom interaction	Welsh (second language) lessons in English-medium schools/colleges	0.75%	30,000
Further Education (FE) classroom interaction	Welsh-medium lessons in a variety of subjects, e.g. Basic skills, Cultural studies, Health care, IT	1%	40,000
Higher Education (HE) classroom interaction	Welsh-medium HE lessons in a variety of subjects, e.g. Linguistics, Arts, Medicine	0.5%	20,000
Adult classroom interaction	Welsh-medium community education	0.5%	20,000
	Adult Welsh lessons	0.75%	30,000

Socialising		22.5%	900,000
Conversations with friends/family	Conversations in the home with groups of friends	7%	280,000
	Conversations with friends/family in public places	6.75%	270,000
	Conversations with friends during leisure activities	6.75%	270,000
Informal interaction at the National Eisteddfod	Conversations in social spaces such as the Merched y Wawr tent, Maes D, the competitors' lounge, y Lle Celf	1.25%	50,000
Informal interaction at the Urdd Eisteddfod	Conversations in social spaces such as the food area, Caffi Mistar Urdd	0.75%	30,000

Private		22.5%	900,000
Conversations with family	Parent(s)-child(ren) conversations	4.25%	170,000
	Conversations with spouses and partners	4.25%	170,000
	Conversations between other family members	4%	160,000
	Phone calls to family members	2%	80,000
Conversations with friends	Conversations with housemates	3%	120,000
	Conversations with close friends in private settings	3%	120,000
	Phone calls to close friends	2%	80,000
		100%	**4,000,000**

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 Steve Morris, who is a Co-Investigator on the CorCenCC project, based at Swansea University.

Profile: Steve Morris

One of the strengths of CorCenCC is that all of us have come to it from a different direction. My background for over thirty years has been in the field of Welsh for Adults [WfA], teaching my first Ulpan [intensive] course in 1982 here in Swansea and returning to my *alma mater* as a lecturer in Continuing Education (Welsh Language) in 1991. The driving force for me (as for so many others working in Welsh for Adults) was, and always has been, one of inclusivity: to ensure that as many people as possible can be given the opportunity to become new speakers of Welsh and be a part of the effort to ensure a vibrant and sustainable future for the language. I have always argued that WfA should be a pivotal part of language planning in Wales not only because it addresses the majority of the population who cannot speak Welsh but also because it is an effective way of creating those new speakers needed to help implement the government's language policies and strategies. To that extent, much of my work revolved around what is usually known as 'status planning'.

It was only until relatively later in my career – having specialised in work on motivation and success in WfA – that I came to know the work of colleagues in Applied Linguistics here at Swansea University, in particular Paul Meara and Tess Fitzpatrick. There was always a strong desire in research in WfA that anything we did should speak to the needs of practitioners and learners within our discipline. The '*applied*' nature of this work (before the university 'impact' agenda had taken off) meant that sometimes it was dismissed as somehow not being 'real research'. I remember being at a WfA research conference where Paul had given a paper on vocabulary acquisition, and it was then that we started to discuss areas of collaboration with each other (Paul said: '*Tell me a problem you want to solve, Steve and I'll help you work out an answer...*') and I realised that the '*applied*' element was what we had in common! Through this, we began the work on creating core vocabularies for A1 and A2 levels in Welsh – my original question to Paul was '*How can we be sure we're teaching the vocabulary our learners need in the absence of a comprehensive corpus of Welsh?*' Here I was then – on the cusp of turning into an applied linguist and dabbling with 'corpus planning'... and then, there was Tess Fitzpatrick.

Through the vocabulary work and the Gregynog seminars, Tess and I discussed more and more ideas for working together and it was clear to me that although our disciplines both had so much overlap and potential for working together, many of us in Welsh had spent years ploughing our own furrows and not engaging with these wonderful, enthusiastic, language people in another part of the campus! It was shortly after I had moved from Continuing Education into the Department of Welsh *per se* that Tess introduced me to Dawn (who at the time was working her way through a bucket list of things she wanted to do to celebrate her 30th birthday). The half day meeting with the three of us at Bonaparte's Café Bar in Bristol Temple Meads station on 1st June 2012 was an unlikely location to kick off a national corpus of contemporary Welsh but it worked and the seed of CorCenCC was planted.

This picture of a non-conformist chapel in Swansea might seem a clichéd representation of the language in a newsletter inviting you to meet the team... but take a closer look. The original building used to stand in a different part of the city and was associated with the first ever Welsh language newspaper to be published, '*Seren Gomer*'. It relocated into the building in the picture in 1962 and in 2015 was purchased by the Chinese Christian community in Swansea. The important message here is that the nature of the Welsh speaking community has changed throughout history and continues to change all the time: we should embrace this! It is a sign of a healthy language. The idea of a speaker of Welsh who is monolingual (apart from some very young children) no longer exists and, as the lead for the work package involved in collecting the data which will go into the corpus, it is of paramount importance that we make sure we are gathering examples of contemporary Welsh in a principled manner which are inclusive of all the varieties of the language existing today.

For me, the potential of all this is massively exciting (that is probably the subject of another newsletter article!). I live in a wonderful city in a fantastic location next to the sea and here now in 2017, the Welsh language is still a big part in the lives of many of us who live here (and all over Wales) albeit in different ways and in different domains to those of the past. We now have education through the medium of Welsh, we text, skype and tweet in Welsh, we watch television from a satellite in Welsh and we still actually speak the language face-to-face with each other across the generations.

'Give us your Welsh!'

A famous Welsh poem talks about how many of us in our later life can see clearly those elements which have moulded us. Looking back, I can see that I have often operated at the interface between Welsh and Applied Linguistics without knowing it. Developing, nurturing and making sure this link evolves is one of the intended legacies of CorCenCC and one to which I am passionately committed.

Steve Morris, S.Morris@Swansea.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