CorCenCC Newsletter

Issue 10: March 2017

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Greetings from the PI

Welcome to the tenth edition of the CorCenCC newsletter. This issue marks one full year of work on the CorCenCC project – and what a year it has been! The buzz word for the last two months of the project has been 'launch'. Since our last edition we have launched the CorCenCC crowdsourcing app, the new-look CorCenCC project website (which will eventually host the corpus), and held a public launch of the project at the Pierhead Building in Cardiff. Busy bees indeed! In this edition of the newsletter we will provide write-ups of each of these launches and we will also introduce you to yet another member of the extended CorCenCC family – project Co-Investigator Irena Spasić.

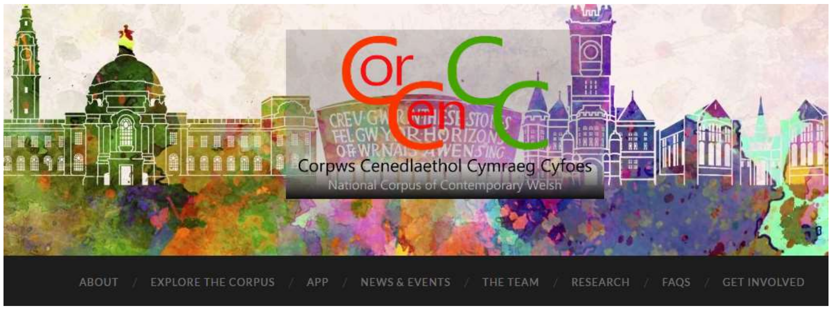
In issues 8 and 9 we provided details of the 'sampling frame' (i.e. data collection plans for the corpus), focusing initially on e-language (issue 8) and then on spoken language (issue 9) as a means of providing a clearer sense of the types/varieties/genres of the data to be included in CorCenCC, and to highlight that we hope to source the data from speakers of all regions, ages and language backgrounds. In this newsletter we finally turn our attention to written language. Hopefully this will provide the final piece of the jigsaw, outlining the 'bigger picture' of the project, and will entice you to get involved and contribute your Welsh!

Happy reading, Dr Dawn Knight (Cardiff University)

News

10/02/2017 – CorCenCC project website launch

The official CorCenCC project website has now been launched! This public-facing site will provide up-to-date information on the progress of the project, details of the vision of the work and the team involved, and will function to answer any questions that interested parties may have about the project. This site will also, eventually, provide the way-in to the corpus, i.e. the site will enable users to access the corpus enquiry tools and pedagogical toolkit. To visit the website go to www.corcencc.org



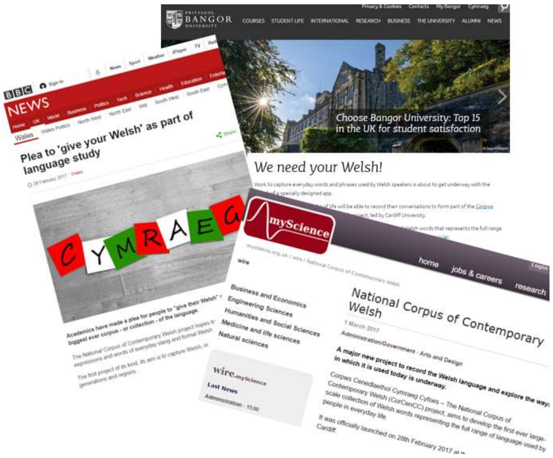
17/02/2017 – CorCenCC Crowdsourcing App launch

To coincide with the launch of the website, February also witnessed the launch of the first release of the CorCenCC crowdsourcing app. The app is currently available on iOS and an Android version will be released within the next two-four months (keep an eye out for that!). The app is easy to use and freely available. To download it now, please use the following QR code or follow this link to the App Store: http://itunes.apple.com/gb/app/ap-torfoli-corcencc/id1199426082

News of the app release was featured on the websites of all partner institutions, on tech websites and in *Y Cymro* and the *Denbighshire Free Press* (amongst others). We are hoping that by spreading the word about the app and project, we can raise people's awareness of the importance and value of the work, and get as many people as possible involved in contributing data and/or using the corpus when it is finally constructed.

28/02/2017 – Project launch

To celebrate a successful first 12 months of the project, the CorCenCC team hosted a launch event at the Pierhead Building in Cardiff Bay. Scaffolded by a weighty media campaign, which included radio interviews on the BBC's *Good Morning Wales* programme (PI Dawn Knight) and BBC Radio Cymru's *Post Cyntaf* (Ambassador Nia Parry) and print and online press coverage in various outlets (including the BBC and Mail Online, institutional websites and tech blogs, amongst others), the event aimed to act as a springboard for engaging with the public, policy makers, educators, publishers and the media; raising awareness about the project and encouraging individuals to support the work.



BBC NEWS

Plea to 'give your Welsh' as part of language study

Academics have made a plea for people to "give their Welsh" for the biggest ever corpus - or collection - of the language.

The National Corpus of Contemporary Welsh project hopes to expressions and words of everyday slang and formal Welsh.

The first project of its kind, its aim is to capture Welsh, in generations and regions.

PRIFYSGOL BANGOR UNIVERSITY

Choose Bangor University: Top 15 in the UK for student satisfaction

We need your Welsh!

Work to capture everyday words and phrases used by Welsh speakers is about to get underway with the a specially designed app.

myScience

National Corpus of Contemporary Welsh

1 March 2017

Administration/Government - Arts and Design

A major new project to record the Welsh language and explore the ways in which it is used today is underway.

Corpws Cenedlaethol Cymraeg Cyfoes – The National Corpus of Contemporary Welsh (CorCenCC) project, aims to develop the first ever large-scale collection of Welsh words representing the full range of language used by people in everyday life.

It was officially launched on 28th February 2017 at Cardiff.

The launch, attended by Alun Davies AM, Minister for Lifelong Learning and Welsh Language, gave guests the chance to find out more about the project, which is a collaboration between Cardiff, Swansea, Lancaster and Bangor universities, and is breaking new ground in creating a large-scale, open access corpus of contemporary Welsh language. Backed by high-profile ambassadors – poet Damian Walford Davies, musician and presenter Cerys Matthews, broadcaster Nia Parry and international rugby referee Nigel Owens – CorCenCC is community-driven and uses mobile and digital technologies to enable public collaboration. A demonstration of our new data collection app which enables Welsh speakers from all walks of life to contribute to the project, was on show at the event. CorCenCC partners and ambassadors also shared their impressions of how the resource will impact on their research, and on the Welsh language community more widely.

Alun Davies, Steve Morris, Dawn Knight, Bethan Jenkins and Tess Fitzpatrick

Minister for Lifelong Learning and the Welsh Language, Alun Davies, said: "I am very pleased to attend the launch of this exciting project today. Not only will this work give us a real record of how Welsh is actually being used, but it will also feed into our aim of developing the role of the Welsh language in technology which will be key if we are to meet our target of a million Welsh speakers by 2050."

Around 85 people attended the launch and the evening also marked the first time that the majority of the extended CorCenCC team were assembled in the same place together! The launch was sponsored by funds from the British Council, the School of English, Communication and Philosophy at Cardiff University, and the Research Institute for Arts and Humanities at Swansea University – many thanks for your support!

The CorCenCC team

01/03/17 – Whole Project Team meeting

Hot on the heels of the launch event, we held the first Whole Project Team meeting at Cardiff University on St David's Day. The meeting, which will take place annually, brings together the CorCenCC Project Team (CPT – which comprises the PI, all CIs, RAs and PhD students), Consultants and all members of the Project Advisory Group, and is a great opportunity for the team to get to know each other a little better (face-to-face) and to discuss ideas and future plans. The aim of the meeting was to provide specific work package (WP) updates, to consider and discuss potential routes to engagement for the project as a whole (concentrating on input mainly from the Project Advisory Group) and to think about how we can best push the boundaries in current corpus research with future developments on CorCenCC.

We would like to say a big thank you to all of you who travelled far and wide to attend this meeting – we all thought it was a very successful and engaging meeting and is likely to provide us with an added strength in ideas and motivation to fuel the next steps of development on the project. We are looking forward to having you all back in Cardiff for the meeting in 2018!

We want your Welsh: A focus on written 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 written resources.

Two of the core issues of corpus design are 'balance' and 'representativeness', which refer to our plans to create a corpus that really reflects the way that Welsh is used in Wales today. In order to achieve this, we need to consider both the authorship and readership of Welsh language written material, as 'use' of the Welsh language includes reception of the language (i.e. hearing and reading it) as well as production. No one single source provides adequate information on which to base a sampling strategy for published material. Sales figures are not necessarily a reliable indicator of readership and ignore the role of lending libraries. Also, there is a paucity of research into Welsh language readership patterns and practices. It therefore seems most appropriate to use publishing data to inform the construction of the sampling frame for written Welsh. In particular, books currently in print give some indication of readership above and beyond actual publication figures and, in the Welsh context, up-to-date and reliable publishing information is readily available. We have therefore used this data as guidance in establishing the sampling frame for published material. In terms of authorship, it is probably beyond the scope of this project to seek balance across authors, especially with regard to location, but we do hope to achieve a gender balance. Where possible, we also hope to record information concerning the authors' linguistic backgrounds as part of the metadata, as this could be of use to future researchers. In addition to published material, we plan to collect written materials from individuals and companies/institutions, etc., again so as to reflect the variety of ways in which the Welsh language is currently used.

Below is the sampling frame for written 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 written language data:**

1. **contact us to find out when the CorCenCC research assistants will be in your area so that they can collect your written language in person**
2. **send us your written language via email** (Microsoft Word/Publisher/Powerpoint documents are ideal, but if you have handwritten language you are welcome to take a photo of it and send us that).

N.B. As this is a corpus of *contemporary* Welsh, we would prefer documents written from 2011 onwards. Thank you!

Medium	Genre	Sub-genre	%	Words
Books			**40.75%**	**1,630,000**
Books for Adults	Fiction	Drama	1.25%	50,000
		Poetry	2.5%	100,000
		Prose	6.25%	250,000
	Eisteddfod	National Eisteddfod compositions and adjudications	1.25%	50,000
	Factual	Humanities, Arts	0.5%	20,000
		Natural Science, Medicine, Social science	0.5%	20,000
		Law, Politics, Education	0.25%	10,000
		Economics, Business, Maths, Accountancy	0.25%	10,000
		Travel, Tourism	0.25%	10,000
		Leisure, Cookery, Puzzles	1%	40,000
		Wales, Local history, Local customs	1%	40,000
		Biographies, Autobiographies	6%	240,000
	Language	Grammar, Language, Dictionaries	0.5%	20,000
	Songs/Hymns	Songs, Hymns	0.5%	20,000
Books for Children	Fiction	Drama	1.25%	50,000
		Poetry	1.25%	50,000
		Prose: secondary	1.75%	70,000
		Prose: primary & nursery	1.25%	50,000
	Eisteddfod (Urdd)	Urdd Eisteddfod compositions	1.25%	50,000
	Factual	Course books	1.25%	50,000
		Factual: secondary	1.25%	50,000
		Factual: primary & nursery	1.25%	50,000
	Language	Grammar, Language, Dictionaries	0.25%	10,000
	Songs/Hymns	Songs, Nursery rhymes, Hymns	0.5%	20,000
Books for Adult Learners	Fiction	Fiction: various	1.25%	50,000
	Factual	Factual: various	1.25%	50,000
		Course books	1.25%	50,000
	Language	Grammar, Language, Dictionaries	0.25%	10,000
Books for Young Learners	Fiction	Fiction: secondary	0.75%	30,000
		Fiction: primary & nursery	0.25%	10,000
	Factual	Course books	0.75%	30,000
		Factual: secondary	0.75%	30,000
		Factual: primary & nursery	0.75%	30,000
	Language	Grammar, Language, Dictionaries	0.25%	10,000

Magazines, Newspapers, Journals		**19.25%**	**770,000**
Papurau Bro	Papurau Bro	10%	400,000
Journals	Academic, Arts, Music, Literature	1.75%	70,000
Society Magazines/Newsletters	Arts, Music, Literature Nature Science Agriculture Women Religion	1.75%	70,000
Newspapers	Monthly	1.25%	50,000
Periodicals	Current affairs: weekly	1.25%	50,000
	Current affairs: monthly	1.25%	50,000
Children's Magazines	Secondary	0.5%	20,000
	Primary	0.5%	20,000
Adult Learners' Magazines	Learners: adults	0.5%	20,000
Young Learners' Magazines	Learners: children (primary and secondary)	0.5%	20,000

Miscellaneous material	**40%**	**1,600,000**
Letters & Diaries	5%	200,000
Formal letters	5%	200,000
Academic Essays, Exam Scripts	5%	200,000
Advertisements	2.5%	100,000
Information leaflets, Political documents, Legal documents	6.25%	250,000
Stories	10%	400,000
Signs	1.5%	60,000
Other	4.75%	190,000
	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 Professor Irena Spasić, who is a Co-Investigator on the CorCenCC project, based at Cardiff University.

Profile: Professor Irena Spasić

We might as well start using Monty Python's catchphrase "and now for something completely different" to introduce members of the team. My career has been rather linear with no sudden U turns. I started studying computer science in high school and never looked back. I graduated in computer science from the School of Mathematics at University of Belgrade in Serbia. This is where I first heard of Noam Chomsky, the father of modern linguistics. Looking back, this idea that we can use maths to model a natural language was probably the first spark that ignited my long-standing interest in natural language processing. During my masters course I continued to be exposed to the ideas of Zellig Harris and his work on sublanguage analysis. Not surprisingly, I picked a topic along these lines for my thesis to explore how we can use a natural language to interface with a database, marrying two of my favourite topics – databases and natural language processing. To continue my research along these lines, I moved to the UK where I did my PhD, again in natural language processing. This is where my approach to natural language processing started shifting from symbolic rule-based approaches to machine learning.

Having graduated from the University of Salford in Greater Manchester, I started a postdoc position at the University of Manchester. I joined an interdisciplinary team lead by Prof. Douglas Kell to develop text mining and database applications in biology. Biology and life sciences in general foster one of the largest text mining communities and have driven research in this area for the past two decades. This was a great opportunity to learn how to communicate with partners who use a variety of their own sublanguages... and how to analyse them automatically (Zellig Harris again!). As I said previously, there have never been any sudden U turns in my career. One thing led to another, and I ended up taking up a lectureship at Cardiff University. What attracted me to this position was the fact that the position was specifically created to build interdisciplinary collaborations with other scientific communities.

The main focus of my academic career has been to establish excellence in research related to text mining, which is the key to gaining knowledge for significant interventions and decision making in the context of big data. This makes it indispensable to other disciplines and has led to deep interdisciplinary collaboration which has been highly successful, leading to impact beyond computer science and informatics as well as developments in my home discipline. I have collaborated with the Schools of Healthcare Sciences, Medicine, Biosciences, Psychology and Social Sciences. So, when Dawn approached me to collaborate on the CorCenCC project, I jumped at the opportunity to work on my two favourite topics – natural language processing and databases. I also fully appreciate the impact such technologies may have in supporting one's national identity. A recent survey by the Pew Research Centre revealed that a language is seen as the most important requisite of national identity. The ability to record language as it evolves by means of crowdsourcing and to use such a corpus to support language acquisition is truly ground breaking and I thoroughly enjoy being part of it.

Irena Spasić, SpasicI@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