CorCenCC Newsletter

Issue 3: June 2016

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Greetings from the PI

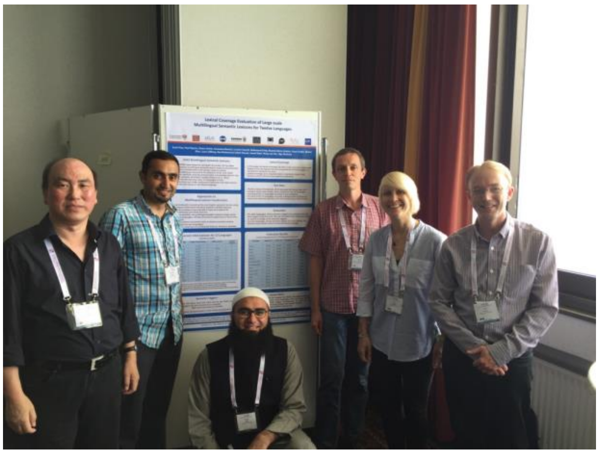
Summer has officially arrived and the conference season has begun. After a little time settling into our roles and fleshing out plans for the project, we are starting to hit the road and will be attending a range of different conferences and events over the summer. The first of these came at the end of May, with Paul, Scott (from WP3) and me attending the LREC (Language Resources and Evaluation Conference) 2016 conference in Portorož. The second was an invited academic seminar talk in Lancaster on 9th June. Brief reports from these events are included in this month's newsletter, along with some details about where you can expect to see members of the CorCenCC team over the summer. In this month's edition we also introduce you to another member of the CorCenCC team - Aran Jones, CEO of SaySomethingin.com and Project Advisory Group member.

Happy reading, Dr Dawn Knight (Cardiff University)

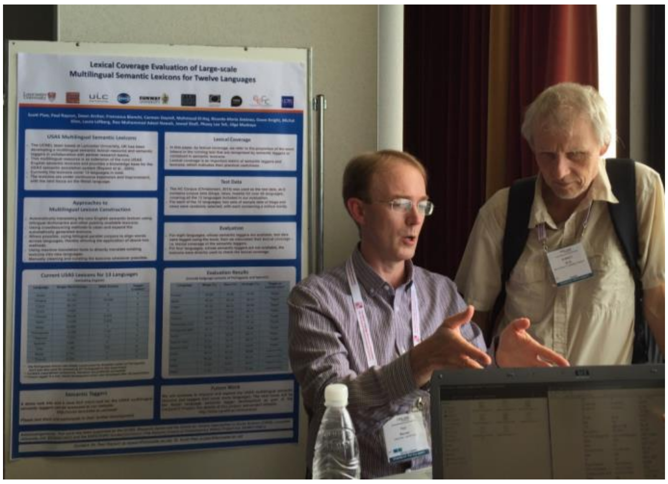
Conferences and events

1. LREC 2016

The 10th biennial Language Resources and Evaluation Conference took place from 23rd to 27th May in Portorož, Slovenia. The Adriatic-Mediterranean coast provided a beautiful backdrop to 5 days of thought-provoking workshops, posters and papers on a variety of different themes; from sentiment analysis and emotion recognition and corpora for language analysis to computer-aided language learning and evaluation methodologies.



Scott Piao, Mahmoud El-Haj, Jawad Shafi, Michal Křen, Dawn Knight and Paul Rayson



Paul Rayson, Scott Piao and Dawn Knight were involved in this two-hour co-authored poster presentation entitled 'Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages'. The poster presentation reported on the construction of large-scale multilingual semantic lexicons for twelve languages, which employ the unified Lancaster semantic taxonomy and provide a multilingual lexical knowledge base for the automatic UCREL semantic annotation system (USAS).

Three other authors, Mahmoud El-Haj, Michal Křen and Jawad Shafi also participated in the presentation, profiling their work on the Arabic, Czech and Urdu semantic lexicons respectively.

The development of the Welsh part of the annotation system is the focus of Work Package 3 (WP3) on the CorCenCC project. A copy of the published paper is here: www.lrec-conf.org/proceedings/lrec2016/summaries/257.html

2. UCREL seminar series

On 9th June project lead Dawn Knight travelled up to Lancaster University to deliver a paper as part of the University Centre for Computer Corpus Research on Language (UCREL) academic seminar series. This invited talk, entitled ‘A community driven approach to linguistic corpus construction’ provided a detailed overview of the plans of the CorCenCC project. The paper discussed some of the practical, methodological and technical issues involved in the design and construction of CorCenCC and sought feedback on the draft outline of the corpus sampling frame that will feature as part of these discussions. It also outlined some of the strategies and techniques that will be employed when we try to engage and recruit participants to contribute data to the corpus, from across Wales.

The presentation was one of three talks delivered during this seminar and was preceded by talks on the spoken and written elements of the BNC14 (British National Corpus 2014) that is currently being developed by Lancaster University, in collaboration with Cambridge University Press. Dawn had an opportunity to talk to the presenters of these talks, Robbie Love and Abi Hawtin to share best practice on corpus planning and construction – a very useful trip indeed (thanks Lancaster!).

Lancaster University
BANGOR
CorCenC
Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh
The National Corpus of Contemporary Welsh: A community-driven approach to linguistic corpus construction
Dawn Knight, Cardiff University
LGC NLW
University of Wales
saysomethinginWelsh

Dawn Knight

3. Other events: June to August 2016

- **IVACS - Inter-varietal Applied Corpus Studies conference 2016 (16-17th June, Bath Spa University)**

The CorCenCC team will present two papers at the biennial IVACS conference – details of these follow (look out for a report on the conference in the next edition of this newsletter):

- **Knight, D., Neale, S., Watkins, G., Spasic, I., Morris, S.** a **Fitzpatrick, T.** Crowdsourcing corpus construction: contextualizing plans for CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes - The National Corpus of Contemporary Welsh).
- **Needs, J., Rees, M., Morris, S., Knight, D.** a **Fitzpatrick, T.** CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – The National Corpus of Contemporary Welsh): Challenges and applications in a minoritised language context.

- **Tafwyl (2nd-3rd July, Cardiff castle)**

Members of the CorCenCC team will be attending the Tafwyl festival, handing out flyers about the project and signing up potential contributors to the corpus. Over the course of the weekend, they will also provide a couple of short presentations to inform members of the public about the main aims of the project, and to provide details of how they can get involved in the work. The first presentation will be in the Literature Wales tent on Saturday 2nd July from 13.00-13.45, to be delivered by Steve Morris and Mair Rees. The second talk will take place in the Cardiff University Yurt – date/time/presenter to be confirmed. See: http://www.tafwyl.org/en/

- **Welsh for Adults conference (8th July, Cardiff)**

Dawn Knight will present a short paper entitled 'Corpora and Pedagogy: developing the community-driven National Corpus of Contemporary Welsh' at the annual Welsh for Adults conference in Cardiff.

- **WISERD conference (13th – 14th July, Swansea University)**

The CorCenCC Management Team, Dawn Knight, Tess Fitzpatrick and Steve Morris will co-present a paper at the annual Wales Institute of Economic and Social Research conference that will be held at Swansea University. WISERD was established in 2008 and 'draws together social science researchers from a number of disciplines including sociology, economics, geography and political science'. More information about this event can be found here: http://www.wiserd.ac.uk/training-events/annual-conference-2016/

- **National Eisteddfod (29th July to 6th August, Abergavenny)**

Members of the CorCenCC team will attend the National Eisteddfod to inform the public about the main aims of the project, and provide details of how they can get involved in the work. All being well, we are also hoping to officially launch the crowdsourcing data collection app at the event. Informal presentations on the project will take place in the Cardiff University tent at 11am on 2nd August and in Swansea University tent at 11am on 3rd August (but you will find us on site throughout both days). For more details, see: https://eisteddfod.wales/

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 Aran Jones, CEO of SaySomethingin.com, who is part of the CorCenCC Project Advisory Group.

Profile: Aran Jones

As a late arrival to the joys of corpora, I was startled when it dawned on me that there was no comprehensive corpus for Welsh - so it's genuinely exciting watching this project develop.

I don't have an academic background in linguistics (I'm well aware that even a UCW Aberystwyth degree in literature doesn't count!). My childhood, though, was spent in Wales, England, Germany, Portugal, Sri Lanka and Malaysia, and my working life began in Zimbabwe and then Dubai - leaving me with a deep-seated interest in languages. I'd always known that I needed to learn Welsh one day - it was my grandparents' language - so I tested my learning abilities on each new language I met.

The results were uniformly disastrous.

I'm still a little surprised that I can now speak Welsh - and a taste of success after so many failures sparked a new interest in methodology, which is how I ended up developing the online, formulaic-language-friendly course SaySomethinginWelsh. We've had about 40,000 people access our content now, in almost 60 countries, and have started work on a small blended learning project with Bangor University - which all feels a very long way from sitting in Zimbabwe for two years failing to learn much more than 'Stay well' and 'Great!' ('Musara zvakanaka' and 'Zvakanaka!', if you're interested, so even claiming two phrases is something of a cheat).

I became circuitously connected to this project when Tess Fitzpatrick and Alison Wray were kind enough to invite me to talk about my work in Cardiff. At that point, I still thought that 'I don't know' was a reasonable answer to the question 'Why does your approach work?' - so Tess and Alison kept patiently inviting me back each year until I found some better answers (which would never have crossed my mind if I hadn't spent all those hours reading papers on linguistics in something of a panic on the train to Cardiff).

I'm watching the project develop with a keen sense of selfishness - it's going to save me a huge amount of work. A great deal of what I do turns upon model dialogues - making them up out of thin air gets a little tiring, and has some natural limitations (anyone who's shared my experience of having a tall, tattooed, bearded, death-metal drummer of a learner complaining loudly that he doesn't want to know how to sail on the bloody sea will know exactly what I mean). A corpus of Welsh, particularly one that will be as strong on spoken Welsh as I'm hoping this will be, shifts us into exciting new realms.

Just for starters, we'll be able to cross-check all our existing material against the corpus, and check for nasty holes or unnecessary inclusions - but beyond that, we'll be able to use it to help guide our approach to critical work on vocabulary extension and listening exercises, and help us to give far better guidance in terms of the choices learners make about dialectical variations.

It will, without question, improve the quality of what we do, and we're already hugely grateful in advance. [Don't stop now! Keep going! Hurry up!].

Aran Jones, aranjones@gmail.com

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 Spasic, 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** - 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