Corpws Cenedlaethol Cymraeg Cyfoes

National Corpus of Contemporary Welsh

Greetings from the PI

Welcome to the May 2018 edition of the CorCenCC project newsletter. The sun is shining, the blossom is out, and the CorCenCC team are limbering up, ready for a summer of data collection and public engagement. We will (hopefully) be coming to a town near you soon! If you are interested in attending future CorCenCC events, or indeed getting involved in data collection or transcription, please get in touch (refer to the end of this newsletter for details). In this month's edition we bring you news of some recent academic conferences that members the team have attended. From Ireland to Japan, we have been working hard to help spread news about the project's aims and objectives, and to update the research community on where we are with the progress on the work. A full write-up on the Welsh Government-funded WordNet Cymru project is also provided, including links to the website – try it now! Last but not least, this month is the turn of project Co-Investigator Enlli Thomas to introduce herself to all of our readers, in our regular 'Meet the Team' slot. This will make you want to dig out your canoe and head to the Taff, I promise!

Happy Reading - Dawn

Contents

+ News and events

IVACS (Symposium (24/02/18)

In February, Mair and I joined hundreds of rugby fans travelling over the Irish Sea to Dublin. But unlike the other passengers on the plane, we weren't going there to see the rugby! The annual IVACS (Inter-Varietal Applied Corpus Studies) symposium was being held in Maynooth, outside Dublin, and we were there to share our experiences of building a Welsh corpus with researchers building Irish corpora.

The theme of the symposium was 'Corpus Research in Challenging Contexts', and it was interesting to learn that the Irish researchers were facing similar challenges to ours. One of the biggest challenges for us is to explain to Welsh speakers that their Welsh is good enough. The CorCenCC project is trying to reflect how Welsh is used by people in their everyday lives. We are <u>not</u> looking for special language – what we want is 'real' Welsh.

Remember – we want your Welsh!

We had a very warm welcome in Maynooth, and we would have liked to stay for more than one night (in accommodation that looked like Hogwarts!), but I'm sure that this was just the beginning in terms of working closely with our friends over the sea.

Dr Jenny Needs

Whole Project Meeting Review (28/03/18)

Some of the CorCenCC project team

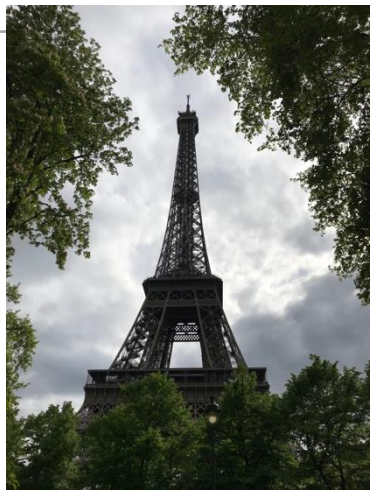
On 28th March, 24 members of the CorCenCC team, from the CMT (CorCenCC Management Team) to the PAG (Project Advisory Group), descended to a sun-filled Cardiff for the second annual CorCenCC Whole Project Team Meeting. The first part of the meeting focused on us saying farewell to members old, welcoming members new, and catching up on developments across Work Packages (WP) over the last twelve months. This included working demonstrations of the CorCenCC part-of-speech (POS) tagger, CyTag (http://cytag.corcencc.org) and of an early demonstrator of the corpus query tool. A more user-friendly version of the latter of these will be on general release in due course – keep checking the website, Twitter/Facebook feeds for more news on this!

Supporting participants from afar

The afternoon focused more on our future users and advisory board members: reflecting on key risks and challenges faced by the project team; how we might engage future users of the corpus and how we can sustain and extend the fabulous work that is being carried out on CorCenCC into the future. I am sure that the team will agree that it was an invaluable and motivating meeting: one which allowed us to congratulate all members of the team on their hard work so far, but also to whet people's appetites on the exciting developments that are soon to come. Watch this space!

Many thanks to all of those who attended the meeting in person and helped to make it such a success – and to those who joined us (bilingually) from afar [see pic].

AFLiCo JET conference (3rd – 4th May 2018)



After navigating various strikes and student mobilisations, I arrived at the Université Paris Nanterre to deliver a plenary presentation at the Journée d'études AFLiCo JET (The French Cognitive Linguistics Association) conference on Corpora and Representativeness. The event attracted a range of speakers/participants from Universities in France and Europe, and a range of topics were presented: from 'the metaphors of death, disease, and sex in a TV show corpus' to reflections on the 'perils and pitfalls of representativeness' in corpus construction. My plenary discussed some of the challenges faced in designing and constructing corpora of minoritized languages, focusing on experiences from the CorCenCC project. It was a very well-organised, friendly event, and my paper attracted a lot of healthy and interesting points for discussion. The wine and cheese reception was a particular highlight too! I even had a sprinkling of time to explore some parts of Paris in the French sunshine, before heading back to Cardiff to prepare for our ensuing trip to Japan (see below). I would like to thank the organising committee for inviting me to participate in this event.

Plenary speakers: Dawn & Thomas Egan (Inland Norway University of Applied Sciences)

Dr Dawn Knight

LREC 2018 (Language Resources Evaluation conference: 7-12/05/2018)

With beaming sunshine and mochi 'aplenty (tasty glutinous rice balls), members of the CorCenCC team travelled to Miyazaki, Japan to participate in the 11th edition of the Language Resources and Evaluation Conference (LREC2018) (7-12 May 2018). This is the largest international language resources conference with 1200+ researchers studying computational linguistics, natural language processing, text mining and digital humanities.

A range of interesting papers/posters were presented at the conference, and the seeds of many future (and hopefully plentiful) collaborations were planted. An enriching and successful time was had by all.

On the first afternoon of the main conference, Scott Piao, Paul Rayson and Dawn Knight presented a poster (co-authored with Gareth Watkins) on the Welsh semantic tagger resulting from work on WP3. The semantic tagger poster described the work on Welsh semantic lexicon construction, challenges presented by the Welsh language. Our evaluation showed that adding the project's Welsh POS tagger (CyTag) helped to improve coverage of the semantic tagger over the Welsh Natural Language Toolkit (WNLT). Paul Rayson said "it was really useful to get feedback and wide exposure for our Welsh taggers in a conference where there is significant interest in under-resourced languages and language varieties".

On the last day, Dawn Knight and Steve Neale presented a poster (co-written with Kevin Donnelly) on CyTag our new Welsh part-of-speech tagger resulting from WP2 (visit http://cytag.corcencc.org for more information and to test the tagger).

The days in between were filled with team bonding, Dawn going for many runs around the golf course, and various explorations of Miyazaki city trying to find a restaurant which served vegetarian food for Dawn and with menus which Mahmoud El-Haj (another traveller from Lancaster) could (hmm!) translate with his mobile phone app from Japanese to English. Social events organised by the conference included a trip to Miyazaki's shrine for the first Emperor of Japan where much sake was consumed.

Paul Rayson, Scott Piao, Steven Neale and Dawn Knight

+ General updates

WordNet Cymru

By Steven Neale

Last month, three of the CorCenCC team members - Irena Spasic, Steven Neale and Dawn Knight - completed work on the WordNet Cymraeg project, which has been underway over a 3-month period in parallel with CorCenCC. WordNet Cymraeg is a lexical database of Welsh content words (nouns, verbs, adjectives and adverbs) grouped together as sets of synonyms, which are then linked to each other according to various lexical and semantic relationships. It follows the same methodology as WordNets in other languages, which have been crucial resources for

determining the meaning of words in natural language processing tasks such as word sense disambiguation and text summarisation.

WordNet Cymraeg has been developed over a 3-month period, for which we were very pleased to be funded by the Welsh Government as part of their Grant Cymraeg 2050 scheme. In line with emerging trends in constructing WordNets automatically, we've leveraged bilingual dictionary information provided by our friends at the GPC (Geiriadur Prifysgol Cymru) to translate words from the English WordNet to Welsh, and then organised those words into Welsh synonym sets based on the original WordNet structure in English. We're really happy with our resulting Welsh WordNet, which covers about 67% of what are considered to be the 'core' synonym sets for a new language - those 5,000 or so concepts that are the most common, and have the most relationships to other synonym sets.

We also had the opportunity to show our work to the funders and to the community as part of the recent Cymru Arloesol event at Tramshed Tech in Cardiff, at which a number of the projects funded by Grant Cymraeg 2050 demonstrated their progress. It was fantastic to be there to see the exciting ways that people are driving the development of Welsh language technology, and for our own Steven Neale to be able to give a presentation on the development of WordNet Cymraeg and the value it offers in that landscape. These are certainly exciting times for the development of technology delivered and available in Welsh, and Welsh natural language processing tools are going to have an important role to play in that.

To find out more about WordNet Cymraeg, visit http://users.cs.cf.ac.uk/I.Spasic/wncy/index.html, or to start using WordNet Cymraeg, files can be found at https://github.com/CorCenCC/wncy.

WP4 update: scope/construct the online pedagogic toolkit

(lead: Enlli Thomas)

Since our last update (September, issue 13), the WP4 team have been consulting Welsh teachers and learners, to see what they think of our plans for the pedagogic toolkit. We have now collected the views of over 40 Welsh teachers and tutors via questionnaires, and have held face-to-face focus groups with 55 teachers/tutors and 14 learners. Once the practical work of creating the toolkit itself has been completed, we will resume the consultation work, demonstrating the tools and getting learner and teacher feedback on them. The technical and creative work has begun on the toolkit's two main elements:

(1) We are very grateful to Anelia Kurteva, a student in Cardiff University's School of Computer Science and Informatics under the supervision of Professor Irena Spasić, who has been busy developing the tools themselves. When designing these, we were inspired by Professor Tom Cobb (consultant on the CorCenCC project)'s excellent tools, LexTutor (https://lextutor.ca/). The focus groups have shown us which tools will be most useful to Welsh teachers and learners, and Anelia has gone about creating these tools and making them attractive and user-friendly. Thanks, Anelia!

(2) In order to show how teachers and learners will be able to use the corpus to learn/teach vocabulary, grammar, etc., we have been writing example materials which will appear on the pedagogic toolkit website. We hope that these will inspire users to create their own materials in future. For example, it will be possible to use the corpus to study noun

gender and the different ways a noun's gender is made obvious by other words in the sentence.

We are really looking forward to demonstrating the first version of the pedagogic toolkit to learners and teachers in a few months' time. But the corpus won't just be useful to learners. What things do you yourself regularly forget and have to check whilst writing in Welsh? Which preposition comes after "falch", perhaps? Or whether a mutation is needed after "rhai"? You will be able to discover the answers to these and to many other questions by looking at sentences from the corpus and seeing how other people have used these words. In the meantime, how about contacting us (corcencc@cardiff.ac.uk) to say how you will use the corpus. What will you look for?

Three Cheers for the Champions!

A very warm welcome to our new crew of champions! Champions are ordinary (but special!) people throughout Wales who have offered to help us collect data by recording conversations with friends, family members and co-workers etc. Amongst them is Ann Lloyd-Biston who lives with her husband, son and daughter in Pontarddulais. Ann works for the NHS but she also has a busy social life. Her big love is singing, and she's a member of two choirs, Parti Llwchwr and the Scarlets' choir. With such a full life, why has Ann decided to be a CorCenCC champion? ' The Welsh language is very close to my heart, it's a core part of my identity. I'm very used to looking after people's health as part of my work, but the health of the Welsh language is very important to me too. I think that CorCenCC is an important project which is going to support the language. By being a champion, I hope that I am contributing in some small way to safeguarding the language and making sure that it's fit and healthy for my children and my children's children. To tell the truth, recording my friends is going to be quite fun!' If you would also like to be a champion, the CorCenCC team would be very happy to hear from you!

+ Meet the team: Enlli Thomas, Co-Investigator, Bangor University

With a name like *Enlli* (Ynys Enlli – Bardsey Island), it is safe to assume that I originate from Pen Llyn (the Llyn Peninsula), right? WRONG! In the haze of the mid 1970s, my parents chose to inflict their second born (not their first – he 'got away' with Delwyn Graham; enough said) with the quaint Welsh tradition of naming your offspring after Welsh places of personal significance, christening me *Enlli Môn* – discernibly prettier than *Bardsey Anglesey*, should the same tradition have existed for English. Now there is nothing wrong with being associated with Pen Llyn, and I do quite like the name Enlli, but there is no geographical or familial significance to it other than my father 'quite liked it' as a name. Whilst I am as proud a north Walian as you will find, and being named after two beautiful islands of north Wales is a quirk that has served me well as an ice-breaker in many situations over the years, I am an even prouder *Mochyn Môn* (literally an 'Anglesey Pig' – harsh, but an actual term of endearment us pigs hold dear to our hearts) – and from Llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch, no less (and

this is all true!) – and so the 'Môn' in my name is actually the only thing that holds any real significance for me.

But, somehow, having a name like *Enlli* or *Enlli Môn* (let's forget the Thomas bit for now) seemed to always make me highly aware of my Welsh language roots and heritage, which triggered my life-long interest in the Welsh language. A constant battle during introductions to strangers around the pronunciation of *Enlli* and the lack of options for using the *to bach* (literally 'little roof') circumflex over 'o' in Môn in anything legal or digital throughout my school years (and to some extent still now) led to a heightened awareness, from an early age, of issues (political, emotional, educational) relating to languages, including a heightened awareness of how language(s) shape the mind and how language(s) shape behaviour.

So, having established where my real interests lay, how was I going to carve out a career in this? I had no real interest in Welsh literature or in creative writing (I studied Welsh at A Level, but preferred the language paper over the literature one) and therefore studying Welsh was not an option for me. (I also studied Music and Mathematics, which I also enjoyed immensely, but I couldn't see myself spending my University years practising the piano and in orchestral rehearsals or dealing with abstract mathematical concepts.) I had never embarked properly on any form of effortful learning of a language (other than dabbling a little with French and German at school), so I felt, at that time, that I had no business studying Linguistics either. So what could I do? Somehow or another, I started thinking about Speech and Language Therapy. Here, I could study language with the core purpose of applying that knowledge to a real situation – and maybe even to make a difference to someone's life. So I went along to UWIC (as it was called then) for an interview. The interview was successful, and I really enjoyed the conversations that we had about Welsh and Welsh-English bilinguals and could see myself getting excited about the prospect of studying this field. However…the campus didn't really do it for me, and, somewhat ill-informed at the time, I felt that the degree limited my options for actual employment (I would be trained to become a Speech and Language Therapist, nothing else). Since there was nowhere else in Wales offering this degree (I was insistent that I would study in Wales), UWIC was an all-or-nothing opportunity. So, for whatever reason, I was put off, and went back to the drawing board.

So, what could I do that would fit my interest but allow for the type of flexibility us Aquarians desire? Psychology, that's what. Psychology was at the cusp of what became a 'psychology revolution' during the 90s with large classes of students and good resources, and allowed me to think about language from the user's perspective whilst also allowing me to continue to study statistics and manipulate numbers. Perfect! And that's what I did. I tailored my course options to suit my interest in language, conducting my undergraduate dissertation in the field of bilingualism from a cognitive perspective, and was extremely fortunate to have two fantastic linguists join the School during my final year, who later supported me through my PhD looking at children's acquisition of grammatical gender in Welsh. And the rest, as they say, is history.

I have had the pleasure of being involved in a variety of funded research studies for over 20 years, spanning psycholinguistic research, to sociolinguistic research, and more recently, to educational research, with one clear theme threading through them all: Welsh. This work has dealt with different bilingual populations (Welsh-English bilinguals, Welsh-German bilinguals, Spanish-Welsh bilinguals; families, children, teenagers, adults; typically developing children and adults; adults with Parkinson's, adults with Alzheimer's, children with literacy difficulties), using a variety of different research methodologies and analyses (qualitative and quantitative), resulting in both academic and practical outputs (standardised vocabulary tests; language monitoring tools for schools) that serve the language and its speakers.

Drawing on my breadth of experience of thinking about the factors that promote and hinder language development, particularly in the minority language context, my involvement with the CorCenCC project is to lead on the development of a pedagogical toolkit that will serve as a useful support for those who are in the process of learning/using Welsh. I am so excited to be involved in this fantastic project, and hope to continue with this collaboration beyond the life of this project.

+ Contact us

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 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 and Jonathan Morris; **RAs** - Steven Neale, Jennifer Needs, Mair Rees, Scott Piao and Lowri Williams; the **PhD students** - Vigneshwaran Muralidaran and Bethan Tovey; **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, 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