CorCenCC Newsletter

Issue 22, March 2019

Corpws Cenedlaethol Cymraeg Cyfoes

National Corpus of Contemporary Welsh

Greetings from the PI

The daffs are out; spring has finally sprung, and the CorCenCC team battle on to achieve the aims that we have been working hard on since back in 2015.

Work is progressing well and as the end draws ever-nearer the fruits of our labour are starting to emerge, like a blossom- enrobed fruit tree in Bute Park. As we begin to release the corpus tools and resources, we would love to gather your feedback. This might include some reflections on our journey so far, your thoughts on CorCenCC going forward, that is, what will you do with the corpus, how might it impact on your own work/practice, and, above all, what CorCenCC (might) mean to you? Check our website and social media accounts for more details in due course! Back to now then: the current issue brings you updates on team changes, some recent project events, and profiles the research of the CorCenCC PhD student Vigneshwaran Muralidaran. Happy reading!

Dr Dawn Knight

Contents

+ News

Farewells…

We are sad to say goodbye to two key members of the CorCenCC team, both of whom have been with us since the start of the project. First, Research Assistant Steve Neale, who has played an important part in all technological aspects of the project, contributing primarily to WPs 2 and 5 (but also working with WPs1, 3 and 4), has recently accepted a position as an AI research engineer at DeepCognito in Manchester. Steve will be working on artificial intelligence and text analytics, beginning his new role in May.

At Swansea, Research Assistant Mair Rees, a member of the WP1 team and a transcriber since the inception of the project, will also be leaving us at the end of March. She'll be taking up the post of trainee translator at the National Assembly of Wales for the first six months, after which she will be a full-time translator with the team there. We hope, therefore, in a way that she may still be indirectly involved with the project as the translation unit there is one of our stakeholders. Mair will be beginning her new post from April.

Both Steve and Mair have been real assets to the project and we wish them pob lwc in their new positions! We hope both will continue to be involved in the CorCenCC project in whatever capacity is possible.

Welcome: Joshua Davies @ Bangor University (WP4)

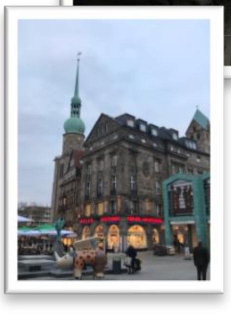
Joshua Davies is a Computing tutor at Bangor International College and a PhD student studying Natural Language Processing at Bangor University. His research involves compression-based language models and their applications. Prior to starting his Ph.D. (supervised by Dr Bill Teahan), Joshua completed a B.Sc. in Computer Science at Bangor University. He also has experience with backend web development for a start-up company in Wales. Joshua is contributing to the development of WP4 by developing and integrating a series of tools for the pedagogical toolkit.

Events: CorCenCC Whole Project Meeting (09/03/19)

On 9th March, 24 members of the CorCenCC team, from the CMT (CorCenCC Management Team) to the PAG (Project Advisory Group), descended on Cardiff for the final 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 a working demonstration of the CorCenCC query tools! The afternoon focused more on our future users and advisory board members: reflecting on the impact and uses that CorCenCC might have, and how we can best engage future users of the corpus. Based on these discussions, we are planning some public engagement and dissemination events for the next few months – so look out for those! It was an invaluable and motivating meeting and, as PI, I would like to send my thanks to all who attended, and to all who have supported the project over the years – we wouldn't be where we are now without you.

Events: IVACS symposium @ Dortmund

Dawn Knight and Steve Morris recently attended the annual IVACS symposium held this year at the Technische Universität Dortmund. It was a chance to update some of the attendees with the latest developments on the project but also to spread the word to those who had never heard of CorCenCC before. The conference theme *Corpus Linguistics and the Classroom* gave us an opportunity to mention our plans within WP4 but also to learn about how similar language communities such as Irish are using corpora to address challenges in the classroom in their own contexts. Many of the delegates commented on how hearing Dawn and Steve's reflections on securing and managing large academic projects had helped and encouraged them in formulating their own future research plans. We look forward to meeting a number of IVACS colleagues at the CL2019 conference later this July.



+ Promoting CorCenCC: St David's Day

The CorCenCC team has been busy promoting the project and encouraging participation on the day of our Patron Saint! Here are some of our St David's Day snaps of the activities in Cardiff.

It's important for the project to be involved in local events in principle and in practice. Academic projects have a duty to provide results that impact a specific community or that contribute to progress in our knowledge of the world around us. To ensure that the project is useful to our 'target community,' we must maintain strong links between project staff and the people who will benefit from the final corpus. This includes participating in cultural events and speaking to members of the public who speak Welsh or who are learning Welsh. For CorCenCC, this means discussing the project with members of the public in order to encourage participation as well as to raise awareness of the project in general in order to promote the final resource.

The Team would like to extend sincere thanks to everyone who has, in the spirit of St David, 'done the little things' (and big things!) by contributing to the corpus. Thanks also to our project champions and to all those who support us. The work of encouraging participation in the project continues, so that the end result is a balanced resource, representative of all kinds of Welsh speakers including people from different regions, of different ages, and of varying levels of Welsh usage in their daily lives. As part of the team's commitment to outreach in the community, local activities and events are vital to attracting contributions of all kinds.

It is important to promote the project so that people will be familiar with the resource once the project is completed. Social media sites like Twitter and Facebook are a great way to achieve this, but the project would not be fulfilling its duty to its target community (anyone who is interested in studying, learning and using the language in any capacity) by relying on social media alone. It is essential to this project and to other, similar projects on language to be supported by and to be used by Welsh speakers whilst having a thorough understanding of their needs. The outreach work will continue until the end of the project in August.

There will be opportunities meet the CorCenCC team and to discuss the project in person at the National Eisteddfod as well as at local events up until the end of August this year.

If you would like to contribute Welsh language data to the project, but are unsure of how to do so, follow our social media accounts https://twitter.com/CorCenCC and https://www.facebook.com/CorCenCC/, download the app or drop us a line at corcencc@cardiff.ac.uk!

+ Life on the inside: CorCenCC work experience

Over the past few weeks we have welcomed Gareth Smith to the CorCenCC team in Cardiff, as he has carried out some work experience on the project. We asked Gareth to provide some reflections on his experiences and some general thoughts on the project as a whole: 'Hi – I'm Gareth and I'm currently studying for an MA in "Welsh and Celtic Studies" at the School of Welsh, Cardiff University. A presentation by members of the CorCenCC project during my first semester sparked my initial interest in their work and the aims of the project.

Having studied various aspects of sociolinguistics in relation to the Welsh language, I realised immediately how essential the proposed language corpus would become for future linguistic research in areas such as dialectology and language use.

One element of the MA course is the "work experience" which offers students the chance to undertake a week of working at an organisation of their choice. It is an excellent opportunity to establish networks, develop skills and sample working life in an area of particular interest. Prior to returning to study, I had spent many years working in I.T. primarily in Business and Technical Analyst roles. I decided, therefore, that joining the CorCenCC team for my work experience would offer me the chance to combine both my language and technology interests.

I have undertaken a variety of tasks during my time with the project including audio transcription, transforming text for database inclusion and ensuring data privacy protocols are observed. It has provided valuable insight into the effort required and challenges faced in developing an efficient, user-friendly language corpus. Although my current placement with the CorCenCC team is nearing an end, I will remain in contact and part of the team through my work as an audio transcriber. And, who knows? I may very well be making use of the corpus in future research…'

Thanks for your hard work Gareth. Your contributions have been invaluable. Good luck with your MA!

+ CorCenCC PhD Research plan: Vigneshwaran Muralidaran, PhD student at Cardiff University

Corpus linguistics displays an empirical bent to linguistic analysis by drawing its insights from a natural language *corpus* - a systematic, planned, structured compilation of naturally occurring texts in a given language. A corpus linguist searches for particular words or phrases in the corpus, the frequency of occurrence of a particular item in a text, and highlight target instances and their context to study different types of linguistic phenomena. Corpora are released with appropriate software interfaces to facilitate these queries through word lists, frequency lists, keywords in a given context and collocated expressions. A corpus might also be further used to develop Natural Language Processing (NLP) tools for languages such as Part of Speech tagging, Syntactic parsing and Semantic analysis.

Syntactic parsing is an NLP module that identifies the grammatical structure of a sentence that can be used further for other information retrieval tasks such as document summarization, sentiment analysis, search engines and machine translation. My goal is to develop a syntactic parser for Welsh as a part of CorCenCC project. Research and development of language technology resources, tools for Welsh are currently at an inceptive stage and a suitable syntactic parser for Welsh language has yet not been developed. Although syntactic parsing itself is a well-explored problem in NLP, there is no single standard framework for grammatical analysis or syntactic annotation because of the differences in theoretical persuasions and goals of researchers. There are various theoretical approaches to grammatical structures in linguistics such as generative grammar, constraint based grammar, structuralist theory, construction grammar, cognitive approaches, quantitative linguistics to name a few. When these theoretical approaches are computationally modelled for natural language processing there is a great diversity of annotation schemes for different languages.

In this context, I am interested in understanding language as a symbolic system with form-meaning pairings right from lexicon to grammar. In my earlier research, my colleague and I identified some interesting functional generalizations in the syntax of Dravidian languages that could be explained well by adopting the theoretical view of usage-based theories like Construction Grammar, Cognitive Grammar. The basic idea is that grammar is not a formal autonomous system that is totally independent of meaning. Every linguistic construction is meaningfully motivated, cognitively construed, schematized through usage, such usage patterns are again meaningfully construed to form new construction schemas. From a finite set of such usage-based construction schemas, infinite set of linguistic expressions can be instantiated and can be assembled to form actual sentences. We also identified discourse relations play a significant role in understanding the construction schemas of Dravidian languages.

Extending this background work, my current research focusses on identifying and applying the relevance of Construction Grammar ideas to automatic learning of grammar from raw Welsh text without any manual annotation. In my current research, I attempt to learn the usage patterns with Artificial Neural Network (ANN) models. An ANN is a computational, non-linear model inspired by the neural structure of human brain that is capable of learning different pattern learning tasks like classification, decision-making, prediction and so on by considering various examples. Some of the challenges involved in the project are: choosing the appropriate architecture for learning, identifying usage patterns with very little or no manual annotation, strategies to evaluate the output patterns learnt by the system.

+ 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, Alex Lovell and Jonathan Morris; **RAs** - Steven Neale, Jennifer Needs, Mair Rees, Ignatius Ezeani and Laura Arman; 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** – Scott Piao, 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