Corpws Cenedlaethol Cymraeg Cyfoes

National Corpus of Contemporary Welsh

Greetings from the PI

So, here we are – the penultimate edition of the CorCenCC newsletter. Welcome! Work on the funded part of the CorCenCC project ends at close of August 2019, but I am pleased to announce that we have been granted a non-funded extension to the work, to the end of May 2020, to ensure that everything is triple checked, tested and ready to launch. We are planning a whole host of public events and roadshows over the next year, so do keep an eye on our social media feeds for details of these! In this edition of the newsletter we update you on some exciting news about project Co-Investigator Enlli Thomas, bring you some updates of recent CorCenCC-related events, and introduce you to another two valuable members of the CorCenCC team. Happy reading – and prepare for a bumper, final, edition in July 2019. The end of the start of CorCenCC is nigh, but the adventure will continue…

Dr Dawn Knight

Contents

+ News and events

CorCenCC CI, Professor Enlli Thomas, Director of Research and Impact, School of Education, Bangor University, has been awarded the Learned Society of Wales Hugh Owen Medal for contributions to educational research, in recognition of her expertise on the Welsh language, bilingualism, and studies into teaching, learning and using Welsh.

The Learned Society of Wales' annual medals were awarded recently at the Royal Welsh College of Music and Drama in Cardiff in a ceremony celebrating achievement in academia. Unfortunately Professor Thomas was unable to attend in person, as she was attending an international research conference.

The medals recognise outstanding contributions in research and scholarship. They are a celebration of the achievements of both the individuals honoured and of the Academic sector of Wales, from universities to schools.

The medals were created to inspire and recognise the long (and often overlooked) legacy of Welsh achievement, while celebrating the exceptional researchers of today.

This year, seven medals, all named in honour of significant figures from Wales' history, were awarded. Accepting the award, Professor Thomas commented: "I'm extremely pleased to have been nominated for this prestigious award this year, and especially grateful to those who nominated me and to the Society for awarding me this wonderful medal and for the recognition of my work. It is a true privilege and an honour to have been able to work in an area that is very close to my heart for over 20 years – the acquisition of Welsh and bilingualism in children – and it is wonderful to be part of the national buzz as we develop strategies and evidence-based educational interventions in order to motivate more users of Welsh by 2050".

CorCenCC @ HealTAC 2019

A panel of researchers, patients, clinicians and companies gathered at the Healthcare Text Analytics Conference (HealTAC) in Cardiff to discuss the state-of-the-art in processing healthcare free text when language barriers exist…When a patient and the health professional who cares for them both have first languages other than English, what's the best way to conduct consultations, keep and share records? As this year's UK Healthcare Text Analytics Conference was organised in the capital city of Wales, the conferences panel groups explored this issue in the case of the Welsh language.

The discussion, held in Park Plaza Hotelon 24–25 April, will support the development of new inspection methodologies, processes and reports to support the implementation of 'More than just words: Follow-on strategic framework', which was published in 2012 by the Deputy Minister for Social Services in the Welsh Government.

The framework aimed to ensure that relevant organisations recognise that language is an intrinsic part of care and that people who need services in Welsh get offered them. This is called the 'Active Offer'. The intention of the follow-on strategic framework 2016 -2019 was to maintain momentum and build on the previous strategy, including practical support for implementing the 'Active Offer'.

The panel explored technical means that can facilitate service provision in Welsh, including terminological resources, software localisation and machine translation, as well as a variety of application areas, such as prescriptions, mental health and allergies.

There were valuable contributions from local researchers, clinicians and international participants with experience in dealing with multiple languages, including Dr Antoine Pironet from the Belgian Cancer Registry who presented their work on processing data from bilingual pathology reports.

The panel was organised by Gareth Morlais from the Welsh Government and the members of the CorCenCC team, Dawn Knight, Laura Arman and Irena Spasić. Cardiff University is part of the CorCenCC project (Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh), which is creating a major language resource for Welsh speakers, Welsh learners, Welsh language researchers and anyone interested in the Welsh language. CorCenCC is a community-driven project, and Welsh speakers from all kinds of backgrounds and of all abilities can get involved.

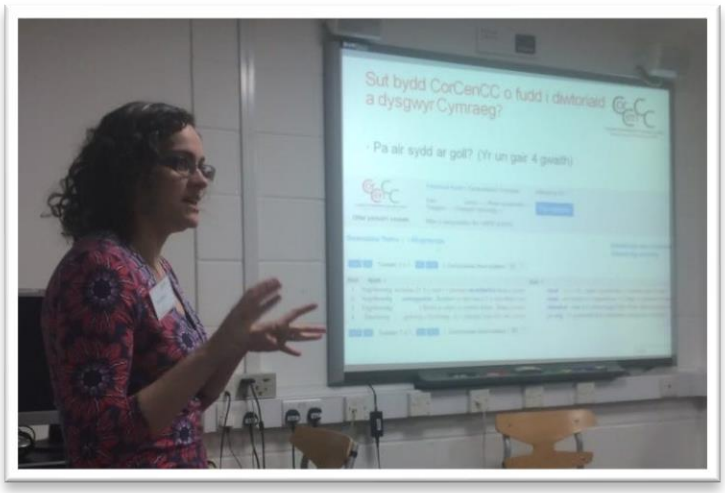
Dawn Knight, from the CorCenCC team, said: "Due to privacy concerns we are currently not collecting healthcare narratives, but CorCenCC's data collection and management infrastructure can be easily deployed in a secure environment to support the development of text mining tools that could facilitate the provision of healthcare services in Welsh."

Irena Spasić, Co-founder of the HealTex network, added: "We have been developing tools and resources that can automate the processing of Welsh text. Our software FlexiTerm can extract domain-specific terminology from Welsh documents on the fly. Our early experiments with parallel corpora demonstrated that domain-specific terminologies can be extracted automatically and mapped between Welsh and English. Such functionality can improve the performance of machine translation of specialised texts such as those found in healthcare."

The conference was sponsored by EPSRC via the UK Healthcare Text Analytics Network, the Connected Health Cities, the SAIL databank and the Data Innovation Research Institute at Cardiff University, as well as Averbis Text Analytics and DeepCongito. The next conference will be organised by Kings College London in 2020.

Conference of the National Centre for Learning Welsh, 11/5/19

The annual conference of the National Centre for Learning Welsh was held in Cardiff Bay on 11 May, and Enlli Thomas and Jenny Needs were there, running workshops with Welsh for Adults tutors. Many thanks to the Centre for the invitation! This was the first opportunity for members of WP4 to be able to demonstrate the work completed so far on the pedagogic toolkit, and to obtain feedback which will lead the rest of the development work.



It was a pleasure to see the positive response of the tutors to the tools and the possibilities the corpus offers to tutors and learners. The corpus has the potential to transform the way that learners come to know the patterns and vocabulary of Welsh, by offering authentic data and allowing learners to discover the answers to their own questions. "Is there a mutation after *rhai*?" … "Which preposition comes after *tueddol*?" … if you ask questions like this, the corpus is for you too!

We're extremely grateful to the tutors who attended our workshops for all the constructive feedback they gave us, as well as their ideas regarding potential future developments. It's essential that the corpus and its associated pedagogic tools meet the needs and expectations of the target users. We want to create resources that are really used and that become an integral part of Welsh lessons across the country.

Welsh for Adults tutors and learners are just one of our target groups. Over the next few months, we will also be meeting primary and secondary school teachers – both Welsh and English medium – in order to ensure that stakeholders in every sphere have the opportunity to influence the creation of the final tools. We look forward to meeting you!

In the meantime, we'd like to ask for your feedback on the proposed name for the pedagogic toolkit – *y Tiwtiadur*. What do you think? Please email corcencc@cardiff.ac.uk with your thoughts.

Y Tiwtiadur

+ Gap filling fieldwork

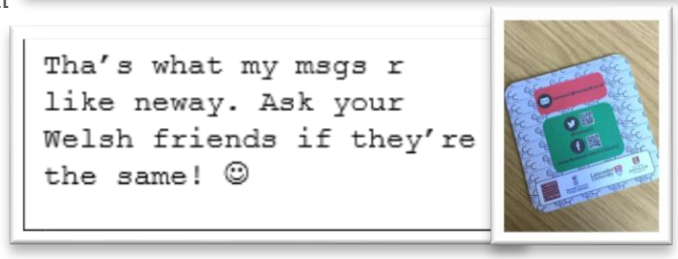
In order to ensure a balanced corpus is created, the WP1 team have worked tirelessly to collect data from each and every corner of Wales for the past 3 years. Now, the time has come to fill the gaps and to concentrate on gathering data from those areas lacking in our word counts, and Denbighshire and Flintshire were first on the list!

The sun was shining on Ruthin and its Mart on Thursday and many locals and visitors contributed their conversations to the project readily whilst catching up with friends and making new ones. We would like to thank Siop Elfair on Clwyd Street for being tremendously helpful and for providing a strategic cuppa! There was a warm welcome to be found in this

part of Denbighshire and plenty of words to add to our targets.

Onwards to Denbigh on Friday where the same enthusiasm was to be found in another part of the county! Denbigh's Language and Play group were especially generous and much data was collected. Thanks to everyone who readily contributed in Denbigh.

Mold on a Friday is all about the livestock auction and it was in a café on site that most of the Flintshire data collection happened, although there was a strong Denbighshire contingent present too! Thanks also to Siop y Siswrn, the Mold Alehouse, and the Saith Seren (Wrexham) for their cooperation and their company for the rest of the day. Adverts for the project were left at the bar, so look out for them on your next visit!



An Easy Way to Contribute: Text messages can often reflect spoken language more closely than other forms of written language. This makes it perhaps the least formal and most contemporary form of written language. CorCenCC is still collecting SMS messages via WhatsApp, so ask Welsh texters you might know to text "helo!" to +447542348512 on WhatsApp for simple instructions on how to select and contribute Welsh SMSs to the project!

+ Data drive

As the CorCenCC project draws to an end, here is another appeal for data which we still need to collect. This appeal went out in the June bulletin of the Association of Welsh Translators and Interpreters but we wanted to share it with all the readers of our newsletter as well. This is a further appeal for different material which we need in order to ensure that the corpus represents all types of contemporary Welsh and to help to reach our 10 million target. The material in which we have a special interest (remembering, of course, that any contributions will be completely anonymised) is:

- Informal letters. Naturally, these are different to text messages and emails (see below) but getting hold of this type of data is quite challenging. If you have any (contemporary) letters of this type that you are willing to share with us, it would be a valuable contribution to the 4 million words of written language that we are currently collecting.
- Adverts / Signs. It's possible that you may have been commissioned to translate specific adverts or signs. Knowing where we could apply for permission to include these (without breaking confidentiality of course) would be extremely useful or perhaps you might know of Welsh language adverts / signs (which haven't been translated) which we might use.
- Emails / Text messages. We have received a number of professional and personal emails to be included in the 2 million digital words in CorCenCC... but there is always room for more! If you would be willing to share any emails (formal or informal), we would be delighted to receive them. Similarly, text or WhatsApp messages. You can contribute these by texting a quick "*helo*" through WhatsApp to +44 7542 348512. You will then receive a link to the permissions form and further instructions.
- Diaries. Once again, getting this kind of data can be challenging. Do you keep a daily diary or a holiday diary, for example? If you have any (contemporary) diaries that you are willing to share with us, we would be really grateful.

If you can send any material (or contribute through the app or consider undertaking transcription work), send us a quick note through corcencc@caerdydd.ac.uk or go to our website www.corcencc.cymru.

+ Meet the team. l

Professor Margaret Deuchar – CorCenCC project consultant

After completing my B.A. in Modern Languages (French and German) at Cambridge and my PhD (on British Sign Language) at Stanford University in California in 1978, my first job was a lectureship at Lancaster University. There I taught mostly sociolinguistics and continued my research on sign language, completing a book entitled *British Sign Language* (published 1984). After a move to Sussex University I became interested in the bilingual acquisition of spoken languages, and got my first ESRC grant for a case study of my daughter acquiring English and Spanish simultaneously. This led to a book co-authored with Suzanne Quay, my PhD student at my next workplace, Cambridge University.

While the book was in progress I was offered a job at Bangor University and arrived there in 1994, coming across Welsh for the first time. As a modern languages student with a Continental orientation through my studies of French and German, I had considered the UK to be a mostly hopeless case of monolingualism! But I was in for a surprise when I arrived in Bangor just before the 1994 Eisteddfod held in Abergele: at this event I was amazed to find that all signs, announcements and events were exclusively in Welsh. At Bangor I signed up enthusiastically for Welsh classes, and continued these until passing my 'A' level in 2000. To prepare for this I volunteered in a Welsh-medium primary school in Colwyn Bay and also joined the local branch of *Merched y Wawr* where I was later elected to Chair, a kind member preparing the script of what I needed to say in meetings.

At that time all teaching at Bangor in my department (Linguistics) was in English, but there were grants available for new Welsh-medium modules. I was successful in getting one of these, and was able to offer new modules to undergraduates in *Ieithyddiaeth Gymraeg* (Welsh Linguistics) by hiring an outside Welsh-speaking lecturer. Students already had the choice of doing their work in either English or Welsh, and a new MA student presented me with a challenge by requesting that all communication with her should be in Welsh. Shortly after this I got my first grant for work on Welsh/English bilingual communication, and recruited a Welsh-speaking RA. With her encouragement, I agreed to hold all our discussions in Welsh, and gradually learned the appropriate linguistic terminology as we discussed theoretical frameworks and the accuracy of transcriptions.

In 2005 I recruited two PhD students with the help of an AHRC grant, to work on Welsh/English code-switching, and one of these students (Peredur Webb-Davies) became a permanent Welsh-medium lecturer in linguistics. In 2007 I was able with colleagues (including Enlli Thomas) to establish the ESRC Centre for Research on Bilingualism with a large grant from the ESRC. In addition to directing the Centre I was able to travel to Patagonia in 2009 to collect Welsh/Spanish data, a fascinating experience. These data, along with Welsh/English and Spanish/English data, are now available at bangor.org.uk, and a co-authored book (with Peredur Webb-Davies and Kevin Donnelly) about the Welsh/English *Siarad* corpus came out in 2018. It was largely because of my work with the *Siarad* corpus that I jumped at the opportunity to act as Consultant in the CorCenCC project after Dawn contacted me in 2012.

Since 2013 I have been based in Cambridge. I am even further from Cardiff than I was at Bangor, and collaboration at a distance has not been easy. However, I have attended all possible CorCenCC meetings and have been impressed with the range of expertise held by the members of the team. I look forward to continuing my association with the project and wish it every success in the future.

+ Meet the team.2

Kevin Donnelly – CorCenCC project consultant

I've been working on free software tools and resources for Welsh and other languages since 2003, including a Welsh version of the leading GNU/Linux desktop, a Welsh-English machine translation system, and Welsh corpora (cymraeg.org.uk). Free software (fsf.org) means that software users have the freedom to run, copy, distribute, study, change and improve the software – the "free" refers to "free" as in "free speech", not as in "free beer" (though most free software is free in that sense too) (gnu.org). The distinction is clearer in Welsh, where the word "free" has different translations for each meaning in English: "rhydd" for the first, and "am ddim" for the second. Most of the internet runs on free software (e.g. the domain name system), and many of the biggest tech companies are built on, and indeed create, free software (IBM, Google, Amazon, Facebook).

I have always believed that for minoritised languages a free license is the only rational choice for tools and resources, because it allows re-use and development – there is no need to re-invent the wheel for each new project, and there is no loss of "community memory" even if a project comes to a halt for lack of funding or other reasons. Sadly, however, it is almost uniformly the case that bodies working on minoritised languages do not share tools and resources freely – they prioritise short-term funding or status over longer-term benefit to the language. This issue is especially pressing in view of the increasing heft of major languages such as English, Spanish, and Chinese in the digital sphere, which makes it more and more difficult for minoritised languages to carve out a space here. The question others will only be able to answer with hindsight in 2050 is: has Welsh missed the boat?

When Dawn was working on getting CorCenCC underway, and contacted me to say that the project would be using an open license, that was the clincher for me, and I was naturally anxious to contribute whatever I could to this major new open resource. In the event, I worked with Steven Neale and others on the tagging sub-project, which drew on two of my previous projects.

One of these was my main Welsh project, Eurfa (eurfa.org.uk), the first (and so far only) Welsh dictionary available under the free software license, the General Public License (gnu.org/licenses/quick-guide-gplv3.html). That has been used for several NLP projects, and the next version, hopefully arriving later this year, should see a major expansion of the dictionary.

The CorCenCC corpora will be available under an open license, but the trail was blazed here by the *Siarad* corpus, assembled at Bangor by Prof. Margaret Deuchar and her team, which used the GPL. I ended up working with Margaret on an automated system to gloss their Patagonia (Welsh-Spanish) and Miami (Spanish-English) corpora (bangortalk.org.uk), and then backported this to handle *Siarad* – the latest version of this (autoglosser.org.uk) was used as a model for Steven's *CyTag*, one of the CorCenCC outputs. In an age of statistics-based NLP, Autoglosser/CyTag takes a rules-based approach, which I tend to think is a better (though more demanding) option for minoritised languages, which may not have a lot of text from which to derive a robust statistical model.

Recently I've been spending less time on Welsh, and going back to my student days at SOAS with two projects. The first is *Andika!,* a tool to typeset classical Swahili poetry in Arabic script, (kevindonnelly.org.uk/swahili). The text can include close and standard transliteration into Roman script, alternate readings, notes, and so on, and allows text from different manuscript versions of the same poem to be tiered to compare wording (e.g kevindonnelly.org.uk/swahili/jaafari).

The second is *Qiezi*, a tool to read Chinese text and typeset it in a number of different formats for teaching or

learning purposes (kevindonnelly.org.uk/qiezi). The version I am currently working on (not yet published) provides lots of information on individual characters, including a mnemonic to assist in learning that character and distinguishing it from others. Knowing 3,000 characters allows you to read 99% of modern Chinese text – so far I've created mnemonics for 1,700 characters, so I'm getting there!

Although the CorCenCC project is winding down, my hope is that it will leave behind a robust set of open tools and resources that will provide a jumping-off point for future NLP work on Welsh (and indeed other languages, since the tools and resource collection procedures can be ported as necessary). It's been a great achievement that the founding CorCenCC project team (Dawn, Tess and Steve M) can be proud of.

+ 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** - Jennifer Needs, 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