Corpws Cenedlaethol Cymraeg Cyfoes

National Corpus of Contemporary Welsh

Greetings from the PI

Happy New Year to all of our readers, and a big welcome to the 21st edition of the CorCenCC newsletter. The project is now in its final furlong, and the final few newsletters will bring you news of roadshows and launches of the corpus – and will demonstrate exactly what you, and others, can do with CorCenCC. In this edition we give you details of a demo version of the query tools that is now live! We invite all readers to test the tools and give us some feedback on what you think. We also introduce you to a new member of the CorCenCC team (and two existing ones, in our 'meet the team' slot), and say farewell to another member who has moved on to pastures new. We also include a plea for more data: while we are getting ever closer to our 10million word target, we would appreciate your help in making it over the finish line – We want your Welsh!

Dawn Knight

Contents

+ News

Goodbye and hello

We are sad to say goodbye to Cardiff University's RA Lowri Williams, who has played an important role in the WP1 team. Lowri has recently taken up the position of Data Scientist for the Data Innovation Accelerator Team in the School of Computer Science and Informatics here at Cardiff University. Lowri has been a real asset to the team and we will really miss her. We wish her pob lwc in her new position, and we hope that she will continue to be involved in the CorCenCC project in whatever capacity is possible. We are pleased to announce that Dr Laura Arman is stepping into Lowri's shoes and has just begun work as our new project RA based at Cardiff University. Laura joins us from Bangor University where she has spent the last two and a half years in Welsh-medium lecturing and pedagogical project work. We asked Laura to say a few words about herself, by way of an introduction to the team…"Happy new year! I'm Laura and I've just joined the project at the beginning of the new year. My background is in linguistics and before getting this post I held various part time lecturing and research jobs at Bangor University, Huddersfield and the University of Manchester (where I got my degrees). I've previously worked on database projects on Kurdish and Romani dialects whilst at Manchester and language contact and minoritized languages are some of my research interests. My PhD, however, was on Welsh morphosyntax and its semantic interface---I'm a Welsh speaker, originally from Bethesda in Gwynedd. Whilst working at Bangor I had the opportunity to do Welsh-medium lecturing in

Lowri Williams

linguistics and this led to my last role as a project officer on the Cyflwyniad i ieithyddiaeth e-book, which is a Welsh-medium introduction to linguistics aimed at first year students. I wrote a few chapters, as well as acting as editor and coordinator for the book which was funded by the Coleg Cymraeg Cenedlaethol and led by Dr Sarah Cooper at Bangor. It will be published in early 2019 by the Coleg Cymraeg. I'm glad to be joining a project which will be a boon to the Welsh language in so many ways! I'm very happy that this resource will be available to learners and researchers alike. I'm also happy to join the team in Cardiff: I enjoy exploring new places and hiking (walking with hiking boots on) and don't know south Wales well at all... yet!''

Welcome to the team Laura! We look forward to working with you on CorCenCC.

Upcoming conference presentations

Members of the CorCenCC team have been accepted to present papers at the following conferences in 2019:

- Knight, D., Morris, S., Fitzpatrick, T., Morris, J., Rayson, P., Spasić, I., Thomas, E.M., Neale, S., Needs, J., Piao, S., Rees, M. and Williams, L. (2019). CorCenCC - Sesiwn arddangos. Paper to be presented at *The XVIth International Congress of Celtic Studies*, July 2019, Bangor University, UK.
- Knight, D. Morris, S. and Fitzpatrick, T. (2019). Peaks and troughs: reflections on securing and managing large academic research projects. Paper to be presented at the *IVACS (Inter-Varietal and Applied Corpus Studies)* one-day symposium on *Corpus Linguistics and the Classroom*, March 2019, TU Dortmund University, Germany.
- Knight, D., Morris, S., Fitzpatrick, T., Morris, J., Rayson, P., Spasić, I., Thomas, E.M., Neale, S., Needs, J., Piao, S., Rees, M. and Williams, L. (2018). CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – National Corpus of Contemporary Welsh): The journey so far. Paper to be presented at the *IVACS (Inter-Varietal and Applied Corpus Studies)* one-day symposium on *Corpus Linguistics and the Classroom*, March 2019, TU Dortmund University, Germany.

Project PI, Dawn Knight, is also leading the organizing committee for the upcoming International Corpus Linguistics Conference, to take place at Cardiff University from 22-26th July. Corpus Linguistics is a biennial conference which has been running since 2001 and has been hosted by Lancaster University, the University of Liverpool, and the University of Birmingham. This will be the first time that the conference will be held in Wales. The theme of the conference is: 'The future of Corpus Linguistics'.

This theme intends to draw our attention to some of the challenges and opportunities which Corpus Linguistics has encountered, continues to encounter and may well encounter in the future. Plans are afoot to hold a CorCenCC event piggybacking on to the conference – more details (including how to get involved) – will be announced soon. For further information on this conference see: www.cl2019.org

+ WP1: Your help needed

The Corpws Cenedlaethol Cymraeg Cyfoes is looking for contributors! Before Christmas, we published our map showing that our spoken Welsh totals were close to 100% for most of Wales, with just a few areas and dialects lagging behind. If you know Welsh speakers from Merthyr Tydfil, Conwy, Flintshire or Anglesey, we need their contributions! They can do this easily and from the comfort of their own homes via our crowdsourcing app (on Android and iOS).

Have a go on the app

Our app allows you to record and send your Welsh in a natural and simple way. We first ask for your permission and your personal details (so that you can tell us to remove your data if you ever change your mind) in the app itself. Then you will be asked to make two recordings. The first is for everyone present to give their consent. The second is the fun part! Maybe you have always liked the way a particular family member speaks and you would like to share their uniqueness with the corpus. Or perhaps you have entertaining conversations with your friends at lunch and you are happy to share a little with us! Whatever the circumstances, we are looking for Welsh language as it is used in everyday, natural settings. Our app is a great way to help us out and to prompt an interesting conversation at the same time!
If you are not too confident with using new apps on your phone, you can follow our detailed instructions here: http://www.corcencc.org/app/

Send us your texts and e-mails

In our corpus of 10,000,000 words, we are looking for a balance of different kinds of language. So far, we have exceeded our targets for language from blogs and websites – happy days!

We are looking for more people to send us their Welsh text messages and e-mails (no matter how many English words they include)! They can be professional or personal and we will anonymize everything before use. If you are keen to contribute to building the Corpus, please forward your **e-mails** to us at CorCenCC@cardiff.ac.uk (and we will need you to sign a consent form for us before we can use anything you contribute). Tell your friends too! We are collecting messages via our WhatsApp. Simply add +44 7542 348512 to your WhatsApp and send us a quick hello! One of the team will then contact you as soon as possible with instructions on how to forward all text messages you are happy for us to include in one quick batch.

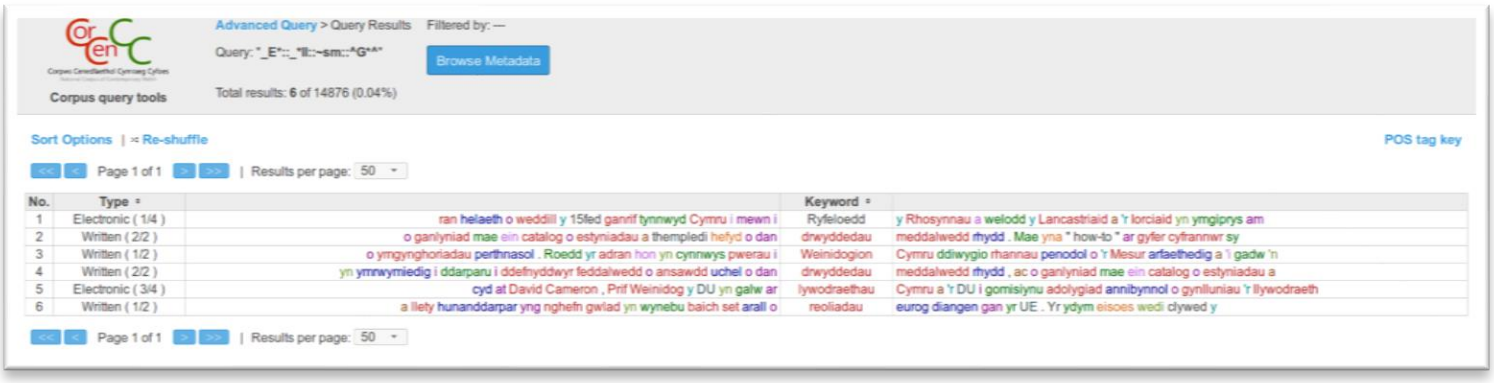
Keep talking

If you would like more information before you share your language data with us, you can talk to us using the same e-mail address or even via Twitter @CorCenCC or our Facebook page CorCenCC. We will be very happy to hear from you!

If you would like to contribute more examples of your speech and text to the corpus, but you do not have time just now, you can follow us on our social media accounts for reminders and updates about the project.

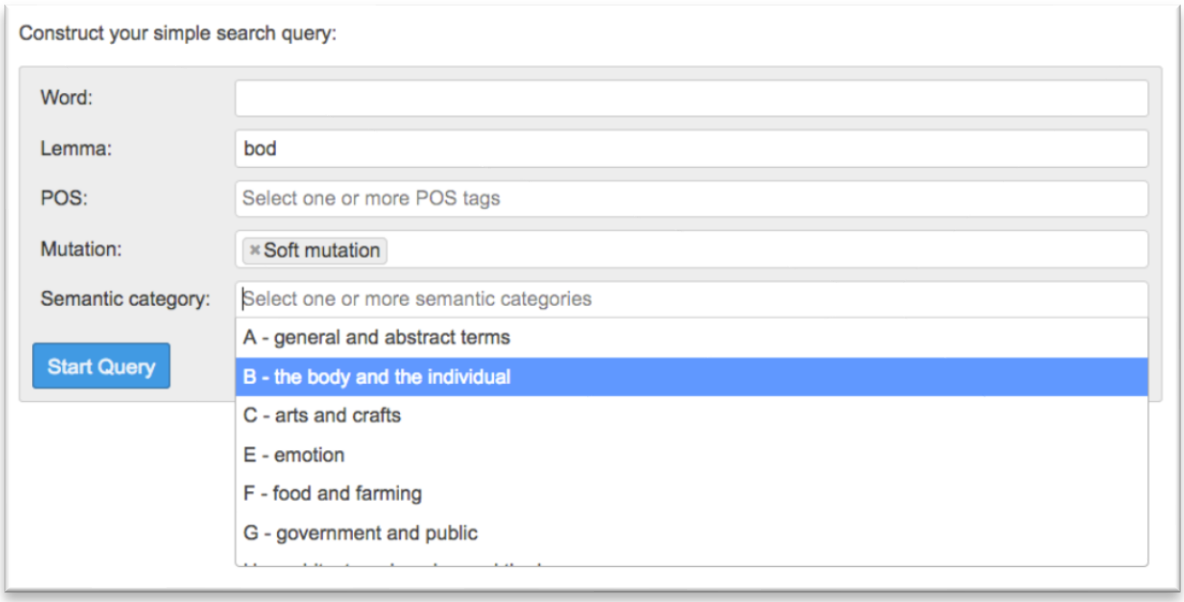
+ CorCenCC query tool demo

The new year has seen some major progress on CorCenCC's front-end corpus query tools, which we are extremely happy to release this month as a beta version. The tools are the major development of WP5, which focuses on the infrastructure required to build and maintain the data and ensure that people can dive into the data when it's ready. Now that we have what we identified as the major functionality up and running, we're hoping that releasing this beta version of the query tools will bring us some vital feedback that we can use to refine and expand the tools' features between now and the end of the project.



The tools currently operate using a very small corpus of around 15,000 words, which we've used throughout CorCenCC to evaluate our various software tools, but of course we'll be replacing this with the data collected by the WP1 team over the next few months. Using the beta version, though, the full range of current functionality can still be accessed and experimented with. This includes:

- Keyword-in-context (KWIC) concordance lines:
    - Search for words (or a sequence of words) in CorCenCC, and see the results in context, with the surrounding text to the left and right displayed. Our query tools offer two ways to produce concordance lines.
        - Select 'Simple Query' > use our handy form to search for individual words, restricted to specific mutation types, parts-of-speech (POS; syntactic categories), and/or semantic categories as appropriate.
        - Select 'Full Query' > use our bespoke query language to chain together more complicated queries and search for sequences of words (full instructions are available on the tools themselves).

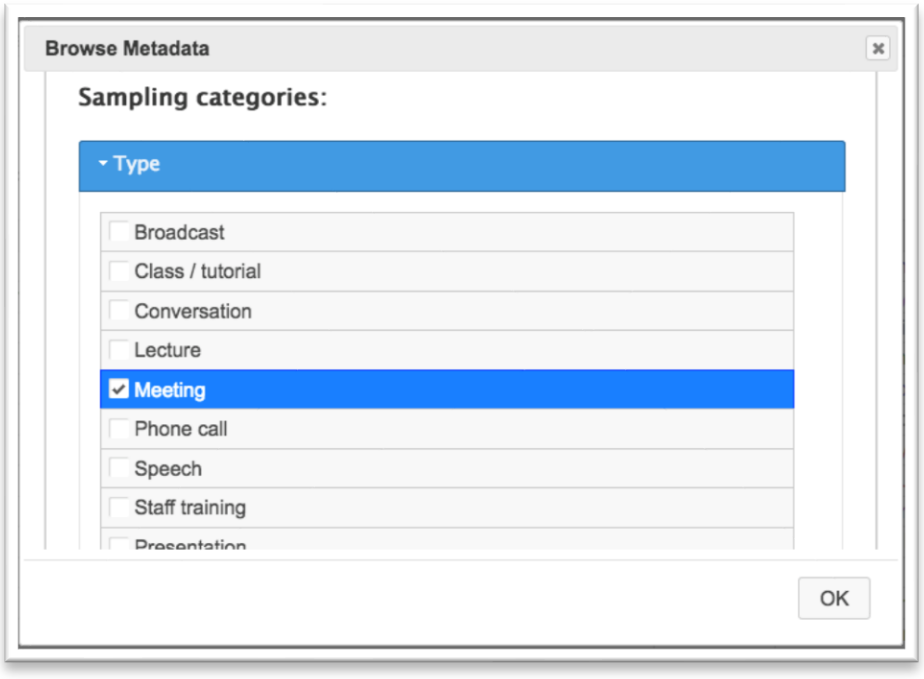


- Frequency lists:
    - Produce lists of the most frequent lexical items in the corpus – whether that be the most common words or the most common lemma forms.
    - Select 'Frequency List' > choose whether to create a list of 'words' or 'lemmas' > choose whether to constrain the list to specific POS tags, mutation types and or semantic tags.

- N-Gram analysis:
  - Produce lists of the most frequent n-grams (also known as 'clusters' of words) – for example, the most commonly occurring sequences of three words used together in the corpus.
    - Select 'N-Gram Analysis' > choose whether to create clusters of 'words', 'lemmas', or 'POS' tags > choose a gram size to produce a list.
- Collocation analysis
  - Produce lists of collocates – words most commonly found occurring alongside a given search term within a given contextual window (n-words either side of the search term) - ranked according to the strength with which they co-locate.
    - Select 'Collocation analysis' > enter a search term (word) for which you want to find collocates > choose a window size within which to consider collocates (plus or minus 7 either side) > choose a 'strength' metric by which to rank collocates (available metrics described on the tools).
- Keyword analysis
  - Produce a list of keywords – those whose occurrence is statistically significant compared to their rate of occurrence in a comparable sub-corpora. For example, find out which words are statistically common in a sub-corpus consisting only of spoken corpus data, compared with a sub-corpus consisting of written and electronic data.
    - Select 'Keyword Analysis' > click 'Sub-corpus A' and select options to include (or not) data for your main corpus (subset of CorCenCC) > click 'Sub-corpus B' and select options to include (or not) data for your reference corpus (subset of CorCenCC) > choose a method by which to rank the 'keyness' of the words (described in more detail on the tools) > select the significance level at which a word is considered 'key'.

These major functionalities are supported by additional features, including being able to:

- Sort results to the left and right of the search word(s), in order to visualise different patterns of Welsh language use;
- Filter results according to the various metadata we've been gathering, about contributions to the corpus and the contributors and speakers involved in them.



In November's newsletter, we described how the development of the query tools has been informed by a survey we conducted on users' preferences in existing corpus analysis and query tools, which has been interesting in helping us decide what functionalities to

prioritise at this stage. In the same spirit, we're also including an option to leave feedback in our beta version of the query tools – so please feel free to tell us what you think of them! We're looking forward to seeing how people make use of the different features we've included, and your input is valuable in helping us decide what features we should include next.

Over the next couple of months, we'll be reviewing all of the feedback we receive on our beta version of the query tools in order to include as much useful functionality as possible, to make the tools as effective as possible for highlighting how Welsh is being used in different contexts across the corpus data we're collecting. Development of our pedagogical toolkit – by members of WP4 – is also well underway, so we'll be working to link that to the query tools so that teachers and learners can utilise the data for their own lesson plans and study sessions.

So, please feel free to explore our beta version of the CorCenCC corpus query tools, currently located at http://corpusdemo.corcencc.org – we look forward to seeing what you all think of it!

+ Meet the team: Mair Parry-Jones, Head of Translation and Reporting Service, National Assembly for Wales

I joined the National Assembly in January 1999 in order to establish the interpretation service. By now I head up the Translation and Reporting Service responsible for four areas of work:

- Text translation;
- Interpretation;
- Producing the Record of Proceedings, and committee transcripts;
- Implementation and monitoring of the Assembly's Official Languages Scheme.

We are a team of almost fifty staff members, some members undertaking more than one of the core duties: some are interpreters and editors; others are text translators and interpreters.

We are all professional language practitioners, using both Welsh and English in our day to day work.

The way we work has changed over the years, mainly as a result of adopting technological innovations.

For years, the Record has been a valuable source of data for Welsh-English parallel corpora. Having such a large corpus of corresponding text in the two languages has been a great help for developments in language technology.

Since 2013, the corpus data is available on a searchable XML version, with an open code for developers to make optimum use of it. Three versions of the Record are now published: a bilingual version, with the contributions appearing in the language spoken in the left-hand column together with a transcript of the interpretation feed from Welsh to English; an English only version, and a Welsh only version.

The style of the Record has also become a more vernacular over the years to reflect the speaker's style of speech in the original language uttered. The style is less formal and more conversational than it used to be.

In addition, since 2017 the Record has been produced with customised transcription software. The TRO system facilitates both the editing and publication to the web processes.

In addition, some team members use Voice Recognition software (or speech to text) to help them transcribe. Unfortunately, this technology is not currently available in Welsh. We hope to see exciting developments in speech technology over the next few years.

On text translation, we worked with Microsoft in 2014 to launch Welsh as a language choice on Microsoft applications on its various platforms, and specifically on Microsoft Office programs. The neural machine

translation system – which introduces an element of artificial intelligence technology to the process – has replaced the statistical system previously used. This has led to a marked increase in the standard of automatic translation.

Many doom-mongers prophesy professional catastrophe as a result of these developments. However, the leading international experts in Artificial Intelligence emphasize that there is – and that there always will remain a need for 'a human in the loop'. The challenge is to adapt to a changing world. Inevitably this will lead to some change, but it will not replace the need for qualified, professional practitioners. The translator's craft has already changed to one of post-editing text produced by a machine. Perhaps the craft should be considered slightly differently – the skill to handle data.

We also have a team of three in-house Welsh tutors who teach Welsh to more than a hundred learners, including many Assembly Members.

Mair Parry-Jones

+ Meet the team

Jennifer Needs, RA at Swansea University

It isn't very difficult to explain why I was eager to be part of CorCenCC as soon as I heard about the project. Languages have always been of great interest to me (I have a degree in Spanish and Linguistics), and Welsh in particular since I completed an MA in the field of endangered languages in 2009. After learning about lesser-used languages across the world, and the steps that can be taken to help stop them disappearing, I decided to come back to Wales to work with the minority language closest to my heart – Welsh.

Originally from Abergavenny, I learned Welsh as a second language in English-medium schools. Welsh was one of my A-level subjects, but I then went to live in England and Spain, and forgot most of my Welsh! I had further lessons through the Welsh for Adults (WfA) scheme and then started working in the field of WfA myself, as a Welsh-medium research assistant. That's a really good way to quickly improve your Welsh!

I went on to do a PhD through the medium of Welsh, again in the field of WfA but this time looking specifically at how online learning can help adults to learn Welsh. I wrote brand new e-learning materials for beginners, following over 100 principles for successful learning. Among these principles were: encouraging learners to "discover" answers themselves rather than relying on the tutor; showing clearly how vocabulary/grammar items are used in context; and allowing learners to control their own learning.

When I heard about CorCenCC, I saw straight away that the corpus would be a unique resource with the potential to enrich the learning experience of thousands of learners of every age, since it combines these important principles. Learners (either independently or with the help of a teacher/tutor) will be able to search for Welsh words and see how these words appear in authentic sentences, exactly as they were spoken/written by the original speaker/author. It will be possible to compare how a word is used differently according to context – e.g. comparing sentences from an academic book with sentences from an informal conversation in the pub. Also, learners will be able to work out grammar rules by looking at example sentences, instead of expecting a tutor to provide the information – e.g. looking at words that come after "hoff" (favourite) and realising that a soft mutation usually follows the word.

By taking a more active role in their learning, Welsh learners should remember new ideas better, and have more fun in class too! I'm really looking forward to seeing CorCenCC's impact on the way people learn and teach Welsh in the future. I'm so glad I was born in a bilingual country and had the privilege of learning Welsh – I hope that CorCenCC will help many others come to know the language better as well. ***Jennifer Needs***

I love an excuse to dress up!

+ 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