CorCenCC Newsletter

Issue 20, November 2018

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Greetings from the PI

Welcome to the 20th edition of the CorCenCC newsletter. In this issue we bring you news of recent CorCenCC- related events and some exciting project developments, including a first look at some of the query tools that end-users will use to browse, search and analyse the corpus. We will also introduce you to another member of the CorCenCC family, Scott Piao, and say a hearty welcome to two new team members, and a sad farewell to Lowri Williams who has left the team. Finally, as this is the final edition of the year, I wanted to take the opportunity to wish all of our readers and supporters a VERY Merry Christmas and all the best for 2019. I look forward to seeing you all again in January, Iechyd da!

Happy Reading – Dr Dawn Knight

Contents

Events

Visit to Bangor Uni (1/11/18)

On a recent visit to team members in Bangor, Dawn Knight, Paul Rayson and Steve Morris met with Delyth Prys, head of the Language Technologies Unit and Gruff Prys, Senior Terminologist at Canolfan Bedwyr. This was a great opportunity to learn more about the cutting-edge work being done at Canolfan Bedwyr in diverse areas ranging from terminology standardization to the well-known on-line dictionary *Ap Geiriaduron* and the development of text and speech technologies (including the amazing *Lleisiwr* project which means that patients who may be at danger of losing their voices are able to store their voice for use at a later stage as a personal digital synthetic one). Readers can find out more about the many areas in which Canolfan Bedwyr is involved through the three on-line portals which can be accessed through the following:

- The Welsh National Terminology Portal http://termau.cymru/?lang=en
- The Welsh National Corpora Portal http://corpws.cymru/?lang=en
- The Welsh National Language Technologies Portal http://techiaith.cymru/?lang=en

This was also an opportunity for the Language Technologies Unit to learn more about CorCenCC and explore ways in which there might be synergy between our work in the future. During the visit we also caught up with other Bangor based colleagues including Kevin Donnelly, Llion Jones (Director of Canolfan Bedwyr) and WP4 lead, Enlli Thomas, and met with Manon Jones from the School of Psychology to discuss potential research overlaps and to spread the word about the vision and aims of CorCenCC.

CUROP poster and dissemination event (Nov 2018)

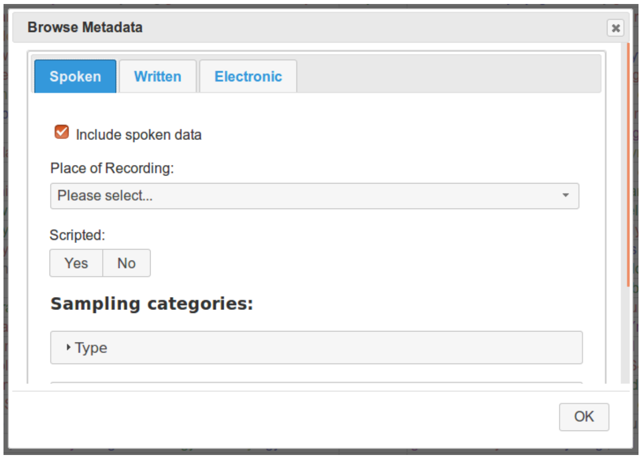
After completing my summer placement working on the WP3 tasks it was then time to prepare for the CUROP (Cardiff University Research Opportunities – a summer research training scheme designed for Undergraduate students) exhibition event, an opportunity to reflect on what I had achieved and present my work. It was a chance to display my contribution to the CorCenCC project through an academic poster (as you can see in the photos) that I had designed myself, and promote the project to my fellow CUROP students. During the event, students and their mentors mingled, informing each other and the public of their various research projects therefore giving us a chance to introduce the project to a wider audience. There was quite an interest in the project and many questions were asked about CorCenCC during the exhibition, all of which I hope I answered well! It was a great way to bring to an end my CUROP placement, and a worthwhile experience of presenting my work and the project to the public.

By Alys Greene

+ Query tool update

Recent weeks have seen things really starting to come together with CorCenCC's front-end corpus query tools, which will provide the gateway for users to access the final corpus data. The tools are being developed as part 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. This is a particularly exciting aspect of the project's technical development, as we begin to construct the tangible output of CorCenCC and realise the means by which our contemporary Welsh dataset is going to be visualised, queried, and analysed.

Our early work has focused on some of the major functionality associated with corpus analysis and query tools, including keyword-in-context (KWIC) concordance lines, frequency lists, n-gram analysis, and collocation analysis. Of course, our principled approach to data collection means that query results can also be filtered according to the various metadata we've been gathering as part of our data collection, which will undoubtedly bring to light some intriguing insights into how Welsh is being used in different contexts! Naturally, the development of the tools is being informed by a recent survey we conducted on existing corpus analysis and query tools, of which there are a wide variety used by researchers, practitioners and linguists for a range of purposes. The feedback we've received about what works well has been really interesting, and we're looking forward to seeing how people make use of the different features we include!



Browse Metadata

☑ Include spoken data

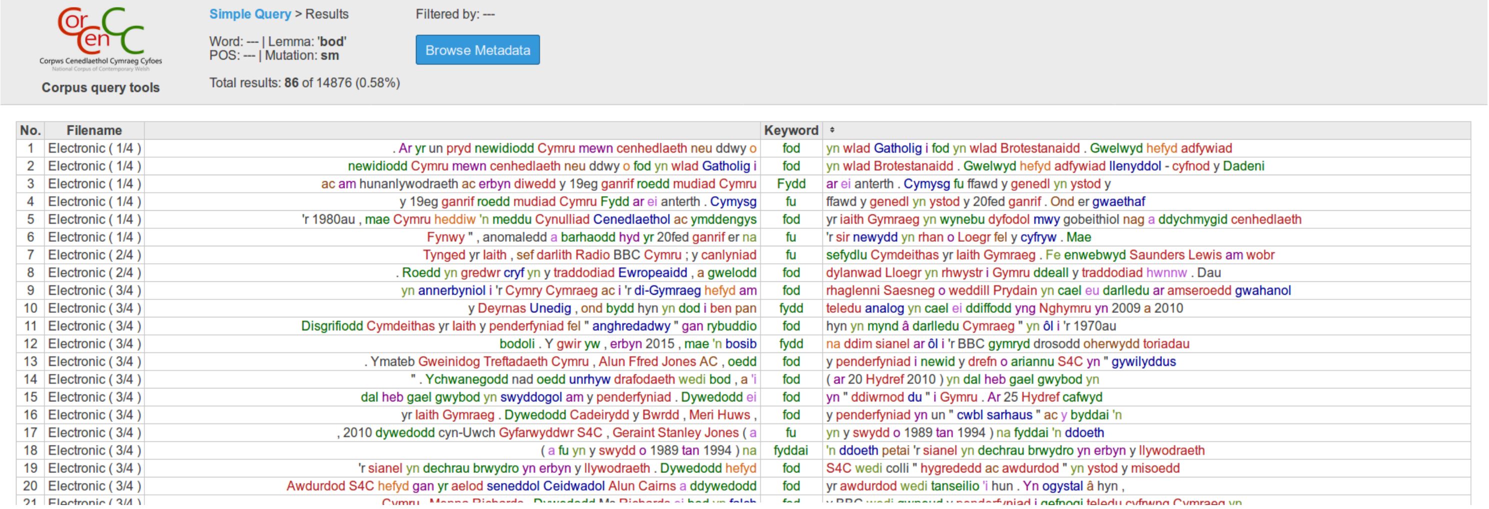
Place of Recording: Please select...

Scripted: Yes No

Sampling categories:

‣ Type

CorCenCC's search tools


CorCenCC
Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh
Corpus query tools
Browse Metadata

Simple Query > Results

Word: --- | Lemma: '**bod**'
POS: --- | Mutation: **sm**

Total results: **86** of 14876 (0.58%)

Filtered by: ---

No.	Filename		Keyword	⇕
1	Electronic (1/4)	. Ar yr un pryd newidiodd Cymru mewn cenhedlaeth neu ddwy o	fod	yn wlad Gatholig i fod yn wlad Brotestanaidd . Gwelwyd hefyd adfywiad
2	Electronic (1/4)	newidiodd Cymru mewn cenhedlaeth neu ddwy o fod yn wlad Gatholig i	fod	yn wlad Brotestanaidd . Gwelwyd hefyd adfywiad llenyddol - cyfnod y Dadeni
3	Electronic (1/4)	ac am hunanlywodraeth ac erbyn diwedd y 19eg ganrif roedd mudiad Cymru	Fydd	ar ei anterth . Cymysg fu ffawd y genedl yn ystod y
4	Electronic (1/4)	y 19eg ganrif roedd mudiad Cymru Fydd ar ei anterth . Cymysg	fu	ffawd y genedl yn ystod y 20fed ganrif . Ond er gwaethaf
5	Electronic (1/4)	'r 1980au , mae Cymru heddiw 'n meddu Cynulliad Cenedlaethol ac ymddengys	fod	yr iaith Gymraeg yn wynebu dyfodol mwy gobeithiol nag a ddychmygid cenhedlaeth
6	Electronic (1/4)	Fynwy " , anomaledd a barhaodd hyd yr 20fed ganrif er na	fu	'r sir newydd yn rhan o Loegr fel y cyfryw . Mae
7	Electronic (2/4)	Tynged yr Iaith , sef darlith Radio BBC Cymru ; y canlyniad	fu	sefydlu Cymdeithas yr Iaith Gymraeg . Fe enwebwyd Saunders Lewis am wobr
8	Electronic (2/4)	. Roedd yn gredwr cryf yn y traddodiad Ewropeaidd , a gwelodd	fod	dylanwad Lloegr yn rhwystr i Gymru ddeall y traddodiad hwnnw . Dau
9	Electronic (3/4)	yn annerbyniol i 'r Cymry Cymraeg ac i 'r di-Gymraeg hefyd am	fod	rhaglenni Saesneg o weddill Prydain yn cael eu darlledu ar amseroedd gwahanol
10	Electronic (3/4)	y Deyrnas Unedig , ond bydd hyn yn dod i ben pan	fydd	teledu analog yn cael ei ddiffodd yng Nghymru yn 2009 a 2010
11	Electronic (3/4)	Disgrifiodd Cymdeithas yr Iaith y penderfyniad fel " anghredadwy " gan rybuddio	fod	hyn yn mynd â darlledu Cymraeg " yn ôl i 'r 1970au
12	Electronic (3/4)	bodoli . Y gwir yw , erbyn 2015 , mae 'n bosib	fydd	na ddim sianel ar ôl i 'r BBC gymryd drosodd oherwydd toriadau
13	Electronic (3/4)	. Ymateb Gweinidog Treftadaeth Cymru , Alun Ffred Jones AC , oedd	fod	y penderfyniad i newid y drefn o ariannu S4C yn " gywilyddus
14	Electronic (3/4)	" . Ychwanegodd nad oedd unrhyw drafodaeth wedi bod , a 'i	fod	(ar 20 Hydref 2010) yn dal heb gael gwybod yn
15	Electronic (3/4)	dal heb gael gwybod yn swyddogol am y penderfyniad . Dywedodd ei	fod	yn " ddiwrnod du " i Gymru . Ar 25 Hydref cafwyd
16	Electronic (3/4)	yr Iaith Gymraeg . Dywedodd Cadeirydd y Bwrdd , Meri Huws ,	fod	y penderfyniad yn un " cwbl sarhaus " ac y byddai 'n
17	Electronic (3/4)	, 2010 dywedodd cyn-Uwch Gyfarwyddwr S4C , Geraint Stanley Jones (a	fu	yn y swydd o 1989 tan 1994) na fyddai 'n ddoeth
18	Electronic (3/4)	(a fu yn y swydd o 1989 tan 1994) na	fyddai	'n ddoeth petai 'r sianel yn dechrau brwydro yn erbyn y llywodraeth
19	Electronic (3/4)	'r sianel yn dechrau brwydro yn erbyn y llywodraeth . Dywedodd hefyd	fod	S4C wedi colli " hygrededd ac awdurdod " yn ystod y misoedd
20	Electronic (3/4)	Awdurdod S4C hefyd gan yr aelod seneddol Ceidwadol Alun Cairns a ddywedodd	fod	yr awdurdod wedi tanseilio 'i hun . Yn ogystal â hyn ,
21	Electronic (3/4)	Cymru . Menna Richards . Dywedodd Ms Richards ei bod yn falch	fod	y BBC wedi gwneud y penderfyniad i gefnogi teledu cyfrwng Cymraeg yn

CorCenCC's concordancer

Over the next couple of months, we'll be expanding on the work done so far to include as much useful functionality as possible, so that users can get all the information they need about contemporary Welsh from CorCenCC. We'll also begin planning for the integration of our pedagogical toolkit – being developed as part of WP4 – to enable teachers and learners to make the best possible use of the data for their own lesson plans and study sessions. Perhaps most excitingly, we'll also begin the task of populating the tools with final data collected by the WP1 team, and that's where we'll start seeing the final dataset taking shape! Keep an eye on our Facebook and Twitter feeds for more information.

Team insights: What do we do with data once it's collected?

If you have been following our newsletters, you'll have seen our updates on our data collection progress. CorCenCC aims to form a corpus of 10 million words from spoken, written, and e-language data. We have adopted several techniques to collecting such data, from using automatic approaches such as web-scraping, to attending events, meeting you all and recording your conversations. But once we have this data, what do we do with it?

Once a member of our research team has collected data, it follows several processing steps before it can be inserted into the final corpus that you, the public, will have access to. For spoken data, the first step is to ensure that it is logged. We note where it was recorded, who the participants of the recording are, and ensure that the quality of the recording is clear. The recording is then safely stored on our secure servers. As CorCenCC will be a text corpus, the recording needs to be converted into text. In this case, we have a team of CorCenCC transcribers who produce a written form of your conversations. Thank-you to our transcribers who continue to work hard on this task!

Before transcriptions, written, and e-language data can be included as part of CorCenCC's final corpus, our research team have the important task of checking the quality of such data. This includes ensuring that all personal information has been removed. Once the quality of data has been checked, it is uploaded to our server and is ready to be inserted into the final corpus. Depending on the size of the data, the task of checking data often takes some time to complete. Nevertheless, it is quite exciting to see the final corpus filling up nicely with your Welsh.

WE WANT YOUR WELSH!

Do you use WhatsApp to text in Welsh? We need your help! We would like to include examples of text messages in the corpus – they're rather a unique way of communicating! Could you send us examples of your messages? It's very easy to do. **Send a WhatsApp message to +44 7542 348512** saying whether you use an iPhone or Android phone and we'll explain what to do next.

+ Meet the team: Scott Piao, former RA at Lancaster University, now project advisor

First of all, it has been a great experience for me to participate in the development of the CorCenCC project and to work with the excellent project team to make the great idea become reality. My interest and experience in the analysis of language data and software tool development for such purpose found a perfect playground in this project, and it has been a great joy for me to collaborate with team members to develop tools for Welsh language, one of the major languages I have ever worked with.

My passion in the development of tools and systems for natural language based information analysis links back to my university undergraduate time in China, when I had an opportunity to study two majors, computing and languages. When I first came to contact with computers, it took no time for me to become obsessed with computer programming. At that time, BASIC programming language (antique today) was the main language, and I still can remember the thrill when I first saw my BASIC program printing out a simple calendar on a piece of print paper with perforated edges with holes (another antique today). As time goes by, BASIC becomes C, again Java, Python, and all sorts of fancy new computer languages, but my enthusiasm about software system development has survived until today. Alongside computer, language has been another side of my core interest. As a trained linguist, I have enjoyed digging into systematic knowledge in languages and linguistics, and for some years I lectured linguistics course. When I came to Lancaster University in 1996, I was so excited to find a new world in UCREL, led by Geoff Leech by that time, where I could combine my skills and knowledge of both languages and computing, and I did a PhD project in Corpus Linguistics in Lancaster University.

Of course, my research interest kept evolving and expanded. My first job in UK in Sheffield University brought me another turn in my research path. In the Natural Language Processing (NLP) Group there, I got an opportunity of learning a lot about NLP area as a complete research field, and grew a strong interest in this area. My knowledge and experience in NLP got another leap during my work in the National Centre for Text Mining (NaCTeM) in the Manchester University. During a series of projects I have worked on over the past eighteen years, now I have developed a wide research interest, spanning NLP, Text Mining, Corpus Linguistics, Social Computing and Data Science, but all is rooted in language data analysis.

Returning to the CorCenCC project, the semantic tagger development for Welsh Language is a continuation of many years' endeavor of developing semantic analysis tools for as many languages as possible based on the English tool initiated by Paul Rayson et al. So far, the Welsh tagger has collected and accumulated language and lexical resources which are the largest among the non-English semantic taggers, thanks to the collaborative efforts of the project team. Of course, due to some unique features of Welsh language and the lack of experience in processing Welsh language in general, it is still a tough challenge to make the Welsh taggers run accurately. In particular, it is difficult to find reliable correlation of syntactic structure between English and Welsh translation of lexical units, and therefore it is highly challenging to fully utilise the existing English semantic resources for building Welsh counterparts. One possibility is to apply deep machine learning techniques to disambiguate semantic categories of Welsh words and phrases based on large corpus data.

From 1st of August, I started a new venture of being an academic lecturer (yes, I am a late starter, hopefully late bloomer, too!), and so have to step down form my senior research associate role in the CorCenCC Project. But I am not going anywhere far, and I will keep close link with this project and make contribution to it wherever possible to help it succeed.

+ Goodbye and hellos

We have, sadly, recently had to say farewell to Cardiff University's Administrative Assistant Lowri Williams, who has accepted a permanent role as a statistical officer at the Office for National Statistics. 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 Alys Greene, a current Undergraduate student at Cardiff University who worked as a researcher on a CUROP placement over the summer (see the related report on page 2 of this newsletter), has stepped into the administrative assistant role. Welcome back to the team Alys – it is great to have you working for team CorCenCC once again!

We are also pleased to announce that we have recruited a new research assistant to work on WP3 at Lancaster University, Ignatius Ezeani. Ignatius Ezeani is a Research Associate at the UCREL Research Centre, Lancaster University. His current research interests revolve around, but are not limited to, developing robust frameworks for adapting existing NLP models and techniques for low resource language research. He is particularly interested in such meaning abstractions and semantic relationships as captured by deep embedding models often trained with huge amounts of data from highly resourced languages and how to project same to low resource languages. He is also generally interested in the design and development of machine learning and deep neural models as well as their applications to, not just NLP, but to the broader field of data science. Ignatius is currently looking at efficient methods to improving the accuracy and reliability of the Welsh Semantic Tagger. Welcome aboard Ignatius!

Alys Greene

Lowri Williams

Ignatius Ezeani

+ 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 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** – 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