CorCenCC Newsletter

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Contents

PI's Greeting (and farewell)

This is it – the final newsletter from the CorCenCC project, marking the end of the main funded period for the work. The first 3 ½ years of the project's journey are now over, but, as mentioned in a previous edition, we have until the end of May 2020 to tie up loose ends and finalise the corpus. The plan is to then release CorCenCC at the end of this period. So, while this is the final formal newsletter, we will keep you all updated on progress via Twitter (@CorCenCC), Facebook (https://www.facebook.com/CorCenCC/) and the website (www.corcencc.org). Keep following us and we hope to see you at a project roadshow/event near you soon!

On behalf of the CorCenCC team, I would like to say a huge thank you to everyone who has supported us along with the way – it has been a real pleasure to meet, and work with, you all. From the project co-Investigators to the consultants; research associates to the PhD students; project advisory board members to the undergraduate researchers; ambassadors to the 1000s of people who have contributed their data, your on-going time and on-going support has been astounding. We really couldn't have got to where we are without your help. In this edition of the newsletter, we provide general updates from the project, in addition to detailed reviews and reflections of each of the work packages (from our work package leads). You also get the chance to meet another two members of the extended CorCenCC family in our usual 'meet the team' feature. Happy reading and keep in touch!

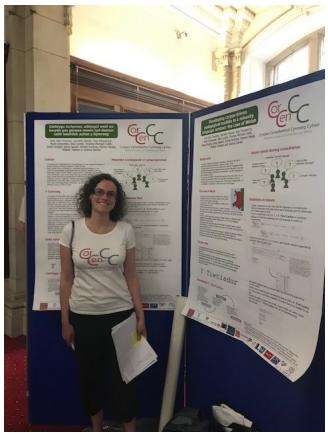
CorCenCC in numbers:

- £1.8m received from the ESRC/AHRC
- 42 months (2015-2019, with a 9 month extension to May 2020)
- 4 academic institutions (1 PI, 7 CIs, 5 RAs, 2 PhD students, 4 UG summer placements, 2 volunteers, 4 PS staff – 3 former members)
- 6 consultants, 12 Project Advisory Group Members, 4 Ambassadors
- 10 formal project meetings, 100s of additional meetings
- 7 mailing lists, 24 newsletters, eyewatering amounts of emails…
- 14 keynotes and over 30 presentations in 11 countries
- 6 publications to date
- 1500+ contributors, circa 100,000 website hits (across www.corcencc.cymru and www.corcencc.org), 1000s of likes and RTs

+ News and Events

CorCenCC @ CL2019, Cardiff (22nd—26th July)

The 10th International Corpus Linguistics Conference (CL2019: www.cl2019.org), organised by project PI Dawn, was held at Cardiff University in the July heatwave. Amongst the 400 corpus linguists in attendance were various members of the CorCenCC team, who delivered 1 workshop, 2 posters and 3 papers based on the project—a conference record! Presentations reported and reflected on developments to date, and provided information on future plans and release of the corpus. The team received some very supportive comments and feedback. A great time was had by all!

CorCenCC @ ACL2019, Florence, Italy

Yes, the weather was lovely, though too hot for some people. There was a lot of funfair, sight-seeing and tasty foods and treats. It was the **57th Annual General Meeting of the Association of Computational Linguistics (ACL2019)**. It was held from the **28th July – 2nd August** at the historical **Fortezza da Basso** in **Florence, Italy** and was a melting pot of new and state-of-the-art ideas and methods for natural language processing and corpus linguistics.

The CorCenCC team, ably represented by Paul and Ignatius, presented our paper on the novel approach to building Welsh part-of-speech and semantic taggers using pre-trained Welsh word embedding models in a multi-task learning scenario. Our new approach exploits deeper linguistic information embedded in the language in assigning part-of-speech and semantic categories using deep neural networks. The paper, titled "***Leveraging Pre-Trained Embeddings for Welsh Taggers***", was presented at the **4th Workshop on Representation Learning for NLP** and has been published in the ACL Ontology.

By Ignatius Ezeani

CorCenCC @National Eisteddfod, Conwy Valley (August)

Despite the adverse weather conditions in the Conwy Valley, the National Eisteddfod at the beginning of August was a roaring success. We would like to thank and congratulate the organisers of the Eisteddfod as well as the Cardiff University Communications team for such a special event.

Steve Morris and Laura Arman presented on the progress of our work as well as on various aspects of the finalized corpus tools in the Societies space and in the Cardiff University tent. People of varied interests and professions attended the talks to hear about the Corpus for the first time as well as to hear how these data might be of use to different stakeholders and the general public.

In the Cardiff University tent, more details were given on the project's design and the data collected. Although the audience was small, potential new contributors showed interest in the project. On the second day of presentations, the pedagogic toolkit was under the spotlight and many tutors and teachers showed enthusiasm for Y Tiwtiadur and its system of generating exercises suited to learners of different levels automatically.

As more of the project's outputs become available, the project will present in different locations so that word gets out to everyone of the Corpus's many advantages to its users.

+ Work package updates and reflections

Work Package 1 (by Steve Morris):

Where to begin?! Time really does fly and looking back over the past three and half years from the WP1 standpoint, it's truly amazing that our wonderful team of three RAs (one being part time) has managed to achieve so much in such a comparatively short period of time. It's easy to forget that when we began back in March 2016, there was no sampling frame, no transcription conventions, no ethics forms, a helluva lot of good will from all our stakeholders/partners but no clear data collection plan and we were all novices to the weird and wonderful world of corpus linguistics (thank goodness that the Lancaster MOOC came out at around the same time!). Of the original team, Swansea RA Jenny Needs (right)

has been with the project right from the beginning. Here, she shares some of her musings about working on WP1: *"A large part of my work on WP1 has involved recording spoken language all over Wales, and looking back at my "field notes", I remember how interesting all the different trips were. I have driven over 3,000 miles, visited 16 of the 22 local authorities in Wales and met over 1,000 Welsh speakers! The first thing I'd like to say is how grateful I am to every single person who agreed to contribute examples of the way they speak to the corpus. Being recorded isn't something that appeals to many people, but I'm so glad you saw the importance of the resource we're building and chose to be part of it. The corpus wouldn't exist without you! I have had the privilege of eavesdropping on an incredible variety of things that happen through the medium of Welsh, including poetry and novel discussion groups, public speaking competitions, adjudications of horse and sheep shows, and public presentations on all sorts of subjects. The best thing in my opinion, though, is that the fieldwork has given me the chance to witness hundreds of you speaking Welsh naturally all across the country, in school or in the workplace, whilst shopping in town, whilst socialising with friends in cafés and pubs, whilst enjoying events such as Eisteddfodau and music and food festivals, and when going about everyday tasks around the house with your families. Keep at it! It's truly fantastic that the corpus will show how evident use of the Welsh language is in every element of the lives of its speakers. Many many thanks to you all for sharing with us an insight into the way the Welsh language is part of your lives. Another thing that has been really heartening, as I travelled the country, was the response from people learning Welsh. It was really important to us that the corpus reflected the fact that new speakers are an essential part of the mosaic that is the Welsh-speaking world. Despite the fact that some of you felt a bit under-confident, you understood how learners would profit from the corpus, and so many of you were willing to contribute to the project for the benefit of future learners. Thank you very much indeed! I would also very much like to thank everyone who is/was part of the CorCenCC transcription team. The work can be challenging at times, I know, but I hope that you too enjoy(ed) hearing examples of all the different ways Welsh is used in Wales today. Without your hard work, all these examples would not be included in the corpus, so you are vital to the success of the project! (No pressure!). I'm really looking forward to seeing the corpus once it's ready. Incredible efforts have been made by everyone who's been involved in the project, and we can be sure that the final resources will be one to be proud of. My own time on the project is coming to an end, but I wish Pob Lwc to everyone who's bringing it all together. It's going to be great!*









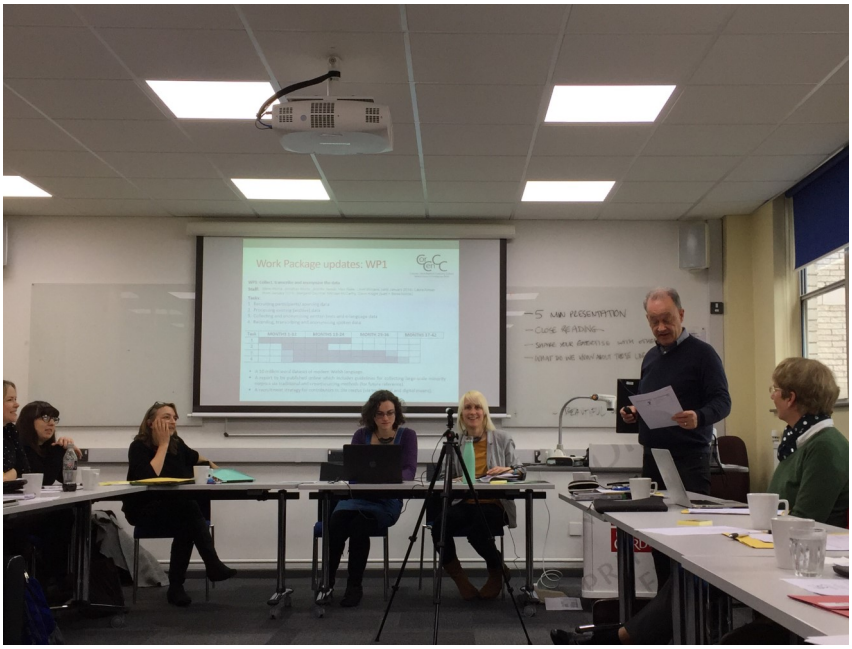
Much of what Jenny has said here is reiterated by the whole WP1 team. Firstly, we should send a massive DIOLCH to all our partners, champions, contributors, participants, transcribers and friends of the Welsh language. It goes without saying that without your support, there would be no CorCenCC. We set ourselves the challenge of collecting a percentage of data from every local authority in Wales in accordance with the number of Welsh speakers in each one in the 2011 census. How closely we have managed to achieve this is very much down to the commitment of

the RAs and their engagement with these local communities. As well as Jenny, we have been fortunate to have an enthusiastic and talented group working in Swansea and Cardiff on WP1, including Mair Rees in Swansea and Gareth Watkins followed by Lowri Williams and finally Laura Arman in Cardiff. You could each tell similar stories to Jenny's of your work with this work package, I'm sure and it is very difficult to pick out highlights. The picture here of Mair with her creation 'Cor-pws' is typical of the innovation and adaptability always evident in the team. 'Cor-pws' came into being as a way of explaining and enhancing the permission sheets for use with our younger age participants. She also made an appearance at the CL2017 conference in Birmingham University – we will make sure she retires to a good home when the project has come to an end!

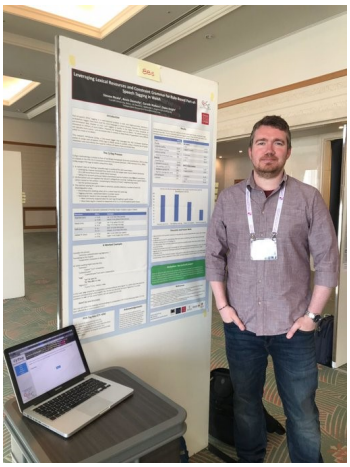
We are now in the strong position of having collected over 90% of the spoken data, nearly 85% of the written data and 190% of the electronic language data. Of course, challenges remain before May 2020:

- We need to add to the spoken and written data (plans are afoot for these);
- We need to work full out on finishing all the transcription / QCing work – **if you know of anyone who might be interested in helping and earning some extra money, please let us know**;
- We need to finalise where CorCenCC will 'reside' and how we will maintain it in the future.

At the recent CL2019 conference in Cardiff, we were able to talk about the work we have done so far as WP1 but also to start to demonstrate CorCenCC to the conference attendees. We followed this up with sessions at the National Eisteddfod in Llanrwst in August and there are plans for roadshows and engagement sessions up until (and probably beyond) May 2020. So, this is not the absolute end. It is, however, a good time to reflect, remember and realise our achievements. It's been a real privilege to lead this fantastic team and your work at the very core of CorCenCC. I think it's also fair to say we've had our fair share of laughs and humour along the way…. some of those stories might have to wait for a future issue though!!!



Pecyn Gwaith 2 (Dawn Knight)



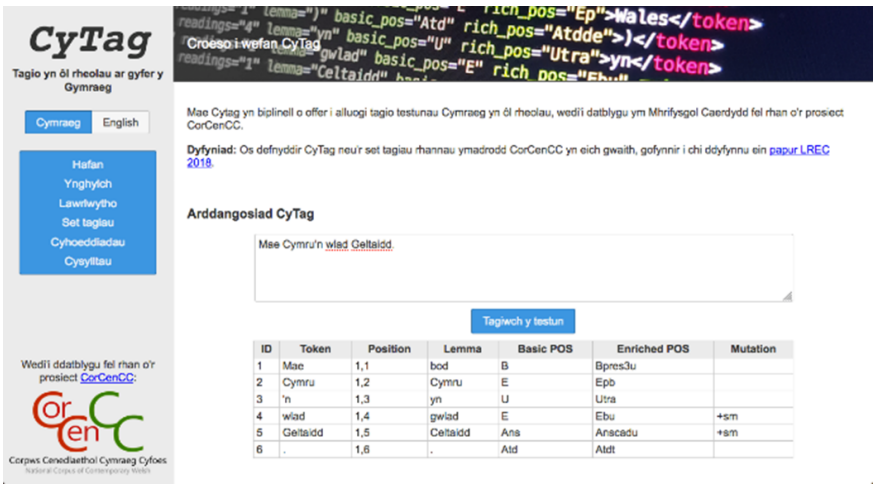
Work on WP2 primarily involved RA Steven Neale and project consultant Kevin Donnelly (led by PI Dawn Knight). The team were tasked to:

- Construct and train a Welsh-language tagger
- Develop an appropriate tagset for the Welsh language
- Tag all data in CorCenCC

Work to date: CorCenCC's bespoke part- of-speech (POS) tagging software, *CyTag*, is complete, robust, and working well. *CyTag* leverages open-source materials to help in the process of deciding on parts-of-speech. It works primarily by using the information found in Kevin Donnelly's *Eurfa* – the largest open- source, freely available dictionary for Welsh – to produce a list of possible tags for each word in a Welsh text. This is supported by specific lists of place names, first names and surnames extracted from Wikipedia data. Once a list of possible words has been produced, we're then able to use a set of bespoke rules to prune the possible tags for each word – based on the tags or features of its neighbouring words – until we arrive at the correct one. For example, in the sentence 'mae Cymru yn wlad Geltaidd' ('Wales is a Celtic country') we can assume that 'yn' is a particle (linking 'Cymru' and 'wlad') because we know that 'wlad' is a soft mutation of 'gwlad' ('country'), and we have a rule to select the particle tag for 'yn' if the following word is a soft- mutated noun. CyTag has been evaluated using a gold standard set of 611 sentences manually annotated with part-of-speech tags for Welsh. The results of the evaluation process were of comparable accuracy to that expected from part-of-speech taggers in other languages (95%+). The website for CyTag was launched in March 2018., see: http://cytag.corcencc.org, CyTag currently contains a text segmenter, a sentence splitter, a tokeniser, and the POS tagger itself.

In addition, CorCenCC part-of-speech (POS) tagset has also long been finalised, and contains 145 'rich' tags, across 13 EAGLES-compliant 'basic' categories. The full tagset is available online here: https://cytag.corcencc.org/tagset?lang=en

A paper, by Steven Neale, Kevin Donnelly, Gareth Watkins and Dawn Knight, describing CyTag and the CorCenCC tagset was accepted for publication at the Language Resources and Evaluation (LREC) conference in Miyazaki, Japan back in May 2018. This presented a thorough technical overview of CyTag and an in-depth evaluation of its accuracy.



Work Package 3 (by Paul Rayson)

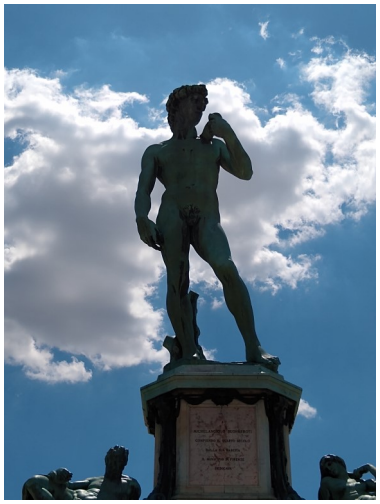
In work package 3 led by Paul Rayson, Scott Piao and Ignatius Ezeani from Lancaster University, our main focus has been to create software and methods for the automatic semantic analysis of Welsh language text. This builds on the tools created in workpackage 2 since part-of-speech tagging greatly helps the semantic level disambiguation. The semantic tagger will be used to tag the whole CorCenCC corpus before the end of the project and the semantic tags will help end users in the corpus query tools for linguistic study and pedagogic tools for teaching purposes. The annotation tools created in CorCenCC are freely available to use, and have also been included in the Wmatrix corpus comparison and annotation system (http://ucrel.lancs.ac.uk/wmatrix/), and can be used to assist with other Welsh language text mining tasks.

Our first task was to develop the semantic tagset for Welsh and this was carried out with the assistance of our project partners in Cardiff and Swansea. Based on and extending the Lancaster USAS system (http://ucrel.lancs.ac.uk/usas/), this includes semantic categories, in the form of tags such as I1.3 (Arian: Prisiau/Money: Price), K5.1 (Chwaraeon/Sports), to apply to Welsh words, idioms and other fixed phrases (also called multiword expressions) based on a semantic classification scheme. This scheme consists of 232 semantic categories falling under 21 major categories, as shown below:

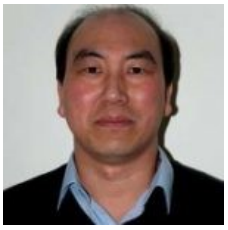
Tag	Definition	Tag	Definition
A	*TERMAU CYFFREDINOL A HANIAETHOL* (GENERAL AND ABSTRACT TERMS)	N	RHIFAU A MESUR (NUMBERS AND MEASUREMENT)
B	*B Y CORFF A'R UNIGOLYN* (THE BODY & THE INDIVIDUAL)	O	SYLWEDDAU, DEFNYDDIAU, GWRTHRYCHAU AC OFFER (SUBSTANCES, MATERIALS, OBJECTS AND EQUIPMENT)
C	*CELF A CHREFFT* (ARTS AND CRAFTS)	P	ADDYSG (EDUCATION)
E	*GWEITHREDIADAU, CYFLYRAU A PHROSESAU EMOSIYNOL* (EMOTIONAL ACTIONS, STATES & PROCESSES)	Q	GWEITHREDIADAU, CYFLYRAU A PHROSESAU IEITHYDDOL (LINGUISTIC ACTIONS, STATES AND PROCESSES)
F	*BWYD A FFERMIO* (FOOD & FARMING)	S	GWEITHREDIADAU, CYFLYRAU A PHROSESAU CYMDEITHASOL (SOCIAL ACTIONS, STATES AND PROCESSES)
G	*LLYWODRAETH A'R CYHOEDD* (GOVENMENT AND THE PUBLIC DOMAI)	T	AMSER (TIME)
H	*PENSAERNÏAETH, ADEILADAU, TAI A'R CARTREF* (ARCHITECTURE, BUILDINGS, HOUSES & THE HOME)	W	Y BYD A'N HAMGYLCHEDD (THE WORLD AND OUR ENVIRONMENT)
I	*ARIAN A MASNACH* (MONEY & COMMERCE)	X	GWEITHREDIADAU, CYFLYRAU A PHROSESAU SEICOLEGOL (PSYCHOLOGICAL ACTIONS, STATES AND PROCESSES)
K	*ADLONIANT, CHWARAEON A GEMAU* (ENTERTAINMENT, SPORTS AND GAMES)	Y	GWYDDONIAETH A THECHNOLEG (SCIENCE AND TECHNOLOGY)
L	*BYWYD A PHETHAU BYW* (LIFE AND LIVING THINGS)	Z	ENWAU A GEIRIAU GRAMADEGOL (NAMES AND GRAMMATICAL WORDS)
M	*SYMUD, LLEOLIAD, TEITHIO A CHLUDIANT* (MOVEMENT, LOCATION, TRAVEL AND TRANSPORT)		

We have taken a number of different approaches to the challenging task of teaching a computer to 'understand' Welsh text, including building lexicons (computer readable dictionaries) manually and automatically, where all words and phrases are classified in the USAS tagset.

We've also prototyped crowdsourcing approaches where non-specialists were asked to classify words, and used machine learning methods to train software based on several hundred high quality test sentences that were manually checked by native Welsh speaking experts on the project. It is a highly challenging task to develop an accurate semantic annotation tool but we have made excellent progress in the three and a half years of the CorCenCC project. Our Welsh semantic lexicon has over 143,000 entries and we've achieved a coverage of over 91% for Welsh running text. Our machine learning tools achieved accuracies around 94-5% for choosing the correct tag in context. In terms of future



plans, we will continue to research improved semantic tagging and using cross-lingual methods to continue extending the system, particularly for multiword expressions which form an important part of the analysis. In addition, we'll be looking to apply the automatic tools to the analysis of other types of Welsh language data. We've demonstrated and presented our research regularly throughout the project including at the UCREL Natural Language Processing summer schools, and at conferences, including:



- Association for Computational Linguistics (ACL), July 2019, Florence, Italy.
- Corpus Linguistics Conference 2019, July 2019, University of Cardiff, UK.
- LREC (Language Resources and Evaluation) 2018 Conference, May 2018, Miyazaki, Japan.
- European Chapter of the Association for Computational Linguistics 2017 (EACL) conference, April, Valencia.

- 1st International Conference on Corpus Analysis in Academic Discourse (CAAD), Valencia, Spain.
- Corpus Linguistics Conference 2017, July 2017, University of Birmingham,
- LREC (Language Resources and Evaluation) 2016 Conference, May 2016, Slovenia.

Work Package 4 (by Enlli Thomas)

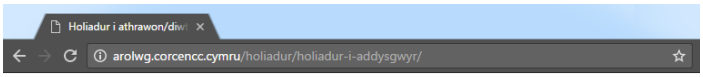
General aims of WP4

One field where corpora can be particularly informative is in language learning/teaching. Corpora can be used to highlight the most common words, phrases and patterns in a language. They can show which words tend to go together, and which ones occur in which types of text (e.g. formal written texts, spoken conversations, professional e-mails, or personal text messages). Corpus users can search for specific words and see them in example sentences. Corpus data therefore provide a rich source of language for the learner that demonstrates how languages are actually used in practice, in various domains and the use of CorCenCC in the language classroom will help demonstrate how Welsh is really used.

Developing Y Tiwtiadur

Following an initial exploration of the kinds of online tool that existed already, including those available to Welsh learners to avoid duplication, the development of Y Tiwtiadur was underway. We were particularly inspired by the work of one of CorCenCC's consultants - Professor Tom Cobb - namely the LexTutor (https://lextutor.ca/), and set about planning for the development of a similar type of tool for Welsh. This development consisted of three main phases: a consultation phase, a product development phase, and a showcasing phase. During the consultation phase, a questionnaire was piloted with a small number of practitioners in the field of Welsh language teaching to explore which resources learners and teachers already use and what they would ideally like to see developed as part of CorCenCC's pedagogical toolkit. Based on their feedback, a new questionnaire was developed for a larger audience. This was distributed to Welsh teachers and tutors at two conferences, and later shared as an online version.

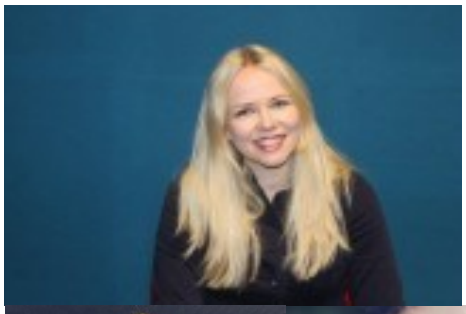
In addition to the questionnaire, we met teachers and tutors face-to-face, to raise awareness of the corpus and the pedagogical toolkit, and to continue to collect views that would help shape the

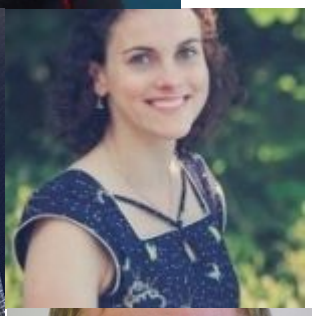


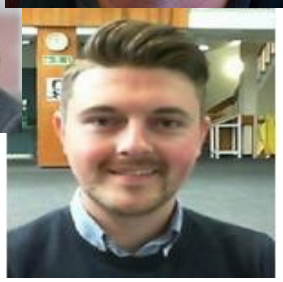
Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Holiadur CorCenCC i athrawon/diwtoriaid Cymraeg
CorCenCC questionnaire for Welsh teachers/tutors

Corpws Cenedlaethol Cymraeg Cyfoes (CorCenCC) yw enw prosiect ymchwil cydweithredol (rhwng Prifysgolion Caerdydd, Abertawe, Bangor a Lancaster) sy'n anelu at adeiladu corpws o'r iaith Gymraeg. Mae corpws yn gasgliad o ddata iaith lafar, ysgrifenedig a digidol sy'n rhoi cipolwg o iaith fel y'i defnyddir mewn sefyllfaoedd 'go iawn'. Fel rhan o brosiect CorCenCC, byddwn ni'n datblygu adnoddau ar gyfer dysgu ac addysgu'r Gymraeg. Mae'n bwysig seilio'n gwaith ar anghenion athrawon a thiwtoriaid, felly **rydyn ni'n gofyn ichi gwblhau'r holiadur byr hwn**. **Dylai fe gymryd tua 20 munud i'w gwblhau**. Mae'r adran gyntaf yn holi am y cyd-destun yr ydych chi'n dysgu ynddo. Yn adran dau rydyn ni'n gofyn am enghreifftiau o adnoddau cyfredol sy'n hynod effeithiol, yn eich profiad chi. Mae adran tri yn holi am y mathau o adnoddau yr hoffech chi eu cael yn y dyfodol. Diolch yn fawr iawn am eich amser. Mae'r holiadur yn dechrau gyda datganiad eich bod yn cydsynio i gyfranogi yn yr arolwg hwn.

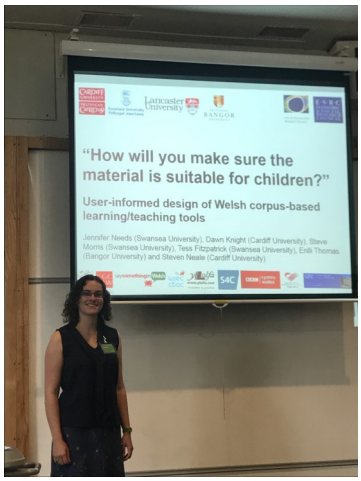




toolkit's development. The questionnaire – along with follow-up focus groups – helped us identify the priorities for CorCenCC's pedagogical toolkit. The discussions we had were extremely useful, and the interchange of ideas that happened over the course of the focus groups and the feedback obtained via the questionnaires enhanced our thinking about how to steer the work as we moved forward. In all, 44 questionnaires were returned by Welsh teachers, trainers, lecturers and tutors from a wide variety of contexts – from those who teach at primary schools (Welsh-medium and English-medium) to those who teach adults – and 10 focus groups were conducted, accessing the views of 55 teachers/tutors and 14 Welsh for Adult learners.

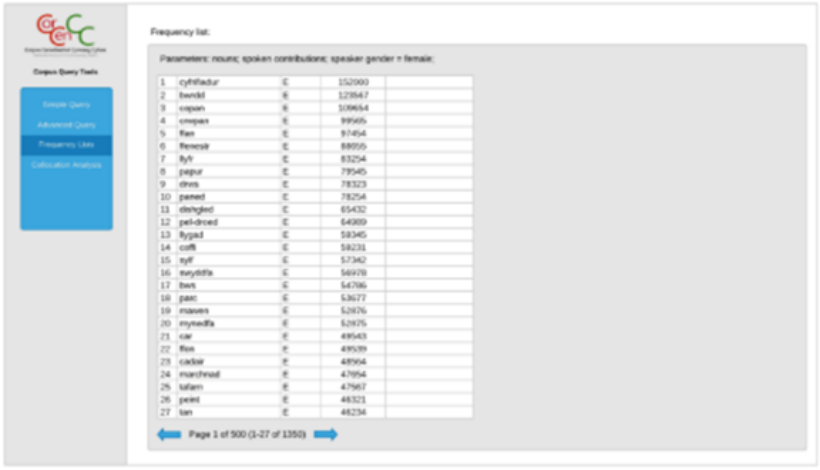
Following the consultation, we worked with members of the WP5 team, initially with Anelia Kurteva, a student in Cardiff University's School of Computer Science and Informatics under the supervision of Professor Irena Spasić, to start working on sample prototypes of the tool, based on the end-user feedback during the consultation. This work was then developed further by Joshua Davies, under the supervision of Bill Teahan, Bangor University, ready for showcasing at the annual conference of the National Centre for Learning Welsh which was held in Cardiff Bay on 11 May, 2019. 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 remainder of the development work. Tutors who attended our workshops provided excellent, supportive and constructive feedback, providing useful ideas regarding potential future developments.

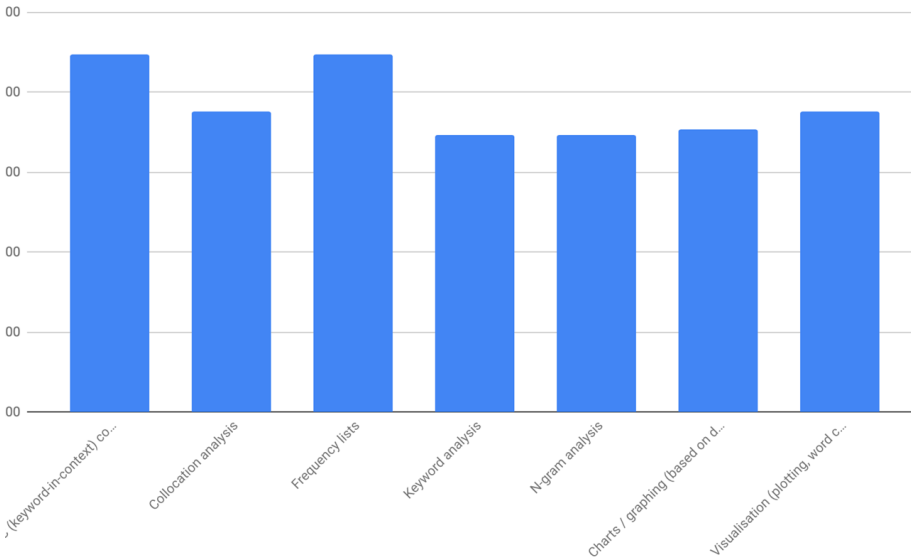


Final plans: Welsh for Adults tutors and learners are just one of our target groups. Over the next few months, we will also be meeting primary and secondary school teachers – both Welsh and English medium – in order to ensure that stakeholders in every sphere have the opportunity to influence the creation of the final tools. We will also liaise directly with Welsh Government officials responsible for the development of the new Curriculum for Wales to ensure visibility of the corpus and Y Tiwtiadur as a potential mode of enhancing engagement with and discovery of the uniqueness of Welsh among L1 and L2 speakers. Our immediate goals include (i) finalising the prototype product, (ii) showcasing the finalised prototype for feedback, (iii) amending the product in line with the feedback, (iv) finalise an on-line guide-book that explains the pedagogical merits of corpus-driven exercises, providing example tasks and exercises for the teacher and learner.

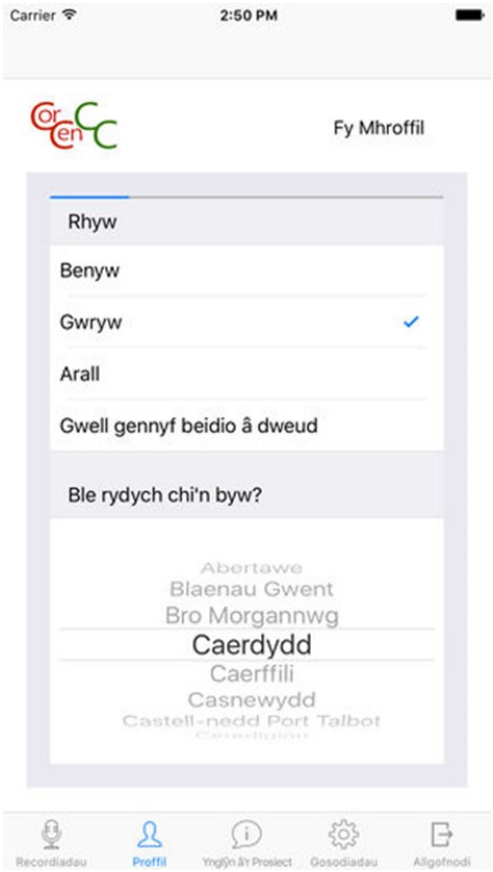
Work Package 5 (by Irena Spasić)

To make the corpus publicly accessible, we developed a bespoke front-end interface to allow users to explore Welsh language data in an interactive, user-friendly manner. We envisage a diverse range of end users including linguistic experts, language learners and teachers, publishers and anyone else interested in studying the Welsh language. To lower the barrier to entry, provide consistent user experience and efficient access to data regardless of the computational platform, we opted for a web-based interface. In that respect, we followed the approach used with other mainstream corpora such as the *BNCweb*, *Corpus of Contemporary American English (COCA)*, the *Michigan Corpus of Academic Spoken English (MICASE)* and the *English Native Edited Japanese Essays (ENEJE)*. Although there are some existing web-based solutions readily available for hosting corpus data online – namely *CQPweb* – we decided to design a bespoke solution for two key reasons. First and foremost, we wanted to tailor its functionality for the Welsh language, whose syntax and semantics we modelled with a language-specific part-of-speech (POS) and semantic tags respectively. In particular, we wanted to enable searching on mutations, a particular characteristic of the Welsh language in which the initial letter(s) of a word change or are omitted depending on the grammatical properties of the preceding word. Further, in an attempt to maximise the utility of the corpus for the purposes of language learning, we needed a modular solution that allows for an easy integration with the pedagogic toolkit also developed as part of this project.





Specific features and functionality have been informed by a survey designed to uncover user preferences from a perspective of a corpus linguists. We wanted to learn what is believed to work well in practice, so that we can comply with best practice and minimise the learning overhead. Equally, we wanted to learn about specific areas that could benefit from innovation. Responses to the survey helped us to focus on the most crucial features for the first development stage and have been extremely useful since then in further refining and expanding on the initial functionality.



Also important in our decision to develop our own corpus interface was the necessity for searches to be able to be filtered based on CorCenCC-specific metadata. CorCenCC's data collection has been principled and largely follows a sampling frame designed to ensure wide coverage of contexts, genres, sources and locations. Thus, it was essential to be able to filter data based on this sampling frame to allow users to study sub-languages and their idiosyncrasies. In addition to collecting existing data, we also actively encouraged collection of spoken data. To facilitate collection of spoken data, we developed a mobile app that allows Welsh speakers to record, annotate and donate their conversations to the corpus in the spirit of citizen science.

We conducted a series of usability studies to evaluate our user-facing tools and continue to improve them in response to user needs and technological advancements. With the support of the Welsh Government, we are already extending the functionality of the corpus in unforeseen directions. We are currently investigating a new way of exploring the Welsh language by navigating through multidimensional space in which words are represented by numbers in a way that reflects the semantic or grammatical relationships between the corresponding words. Other spin off projects include Welsh WordNet (a lexical database of Welsh in which words are grouped into sets of synonyms organised into a network of lexico-semantic relationships: https://users.cs.cf.ac.uk/I.Spasic/wncy/), Welsh FlexiTerm (automatic recognition of multi-word expressions) and Welsh stemmer (reducing inflected and derived words to their root form: https://github.com/ispasic/FlexiTermCymraeg).

Our achievements are the result of team effort. The key members of the team include Dr Steven Neale (a postdoctoral research associate), Prof. Laurence Anthony (a consultant), Prof. Irena Spasić (co-investigator) and Dr Dawn Knight (principal investigator). We have engaged a new generation of developers through student contributions including those of Corey Watson, Anelia Kurteva, David Owen and Vigneshwaran Muralidaran. Our efforts to make our software solutions sustainable have been supported by research software engineers from the Data Innovation Research Institute, Ian Harvey and Dr Jeffrey Morgan. In our venture to study language from a mathematical perspective, we have recently been joined by Dr Padraig Corcoran and Dr Geraint Palmer. Last but not least, we would like to thank the public for taking their time to contribute to the success of the project by means of crowdsourcing.




+ Meet the team

Alex Lovell, CI, Welsh Department, Swansea University

Since my school days, I have been always been interested in languages and the Welsh language in particular. After studying A Level Welsh Second Language at school, I decided to continue studying Welsh at Swansea University in 2010. And I haven't left Swansea since! After completing a BA degree in Welsh in 2014, I started my MA through research degree which was upgraded to Ph.D. in 2015. In my Ph.D., I examined how best the delivery of Welsh as a second language can be successfully supported in the context of English -medium secondary schools. This is an area of research that is very close to my heart, so it was a privilege to be able to give something back to this field. In September 2017, I joined the teaching staff of the Welsh Department in Swansea University, after seven years of study there. I am primarily responsible for teaching on a number of language modules for second language students, but I also co-ordinate and contribute to modules in the fields of language, education, applied linguistics and language planning.

Like all language teachers, I am always looking for new learning resources, not only to improve the standard of teaching but also to improve the experience and quality of learning for the students. This is why I am delighted to be a part of WP4 team of the project – this national corpus has the potential to transform the way we deliver Welsh language teaching and learning in the classroom and beyond. By using the pedagogic tools currently being developed by the WP4 team, teachers will be able to tailor learning and assessment to better suit the needs of their learners. In addition, learners will be able to explore the corpus, encouraging them to take responsibility for their own learning and to develop into independent learners. As someone who often turns to the web and dictionaries to learn how new words (for me) are used in specific contexts, I am truly excited about the educational possibilities that the corpus will offer me and my students in the future.

+ Meet the team

Ignatius Ezeani, RA, Lancaster University

I am new to Natural Language Processing (NLP) and Computational Linguistics (CL). In 2000, I got my first degree in Computer Science from Nnamdi Azikiwe University, Nigeria. Back then, I didn't plan to remain in research or academia so I setup of a small IT support firm in the university town of Awka in Anambra State, Nigeria. However, within the first few years after my graduation that I was literally sucked back into the university where took up a Graduate Assistantship role.

The job put me on the path of an academic career involving teaching computer science modules and supervising student projects. Having found myself back in an academic environment, I enrolled for and obtained an MSc in Advanced Software Engineering from Bournemouth University, UK in 2006. I had to go back to Nigeria after the MSc program to continue with my contract.

My journey with NLP started with the invitation by a colleague, Dr Ikechukwu Onyenwe, who was then a doctoral student at the University of Sheffield. He was pioneering the research work on building language resources (corpora and tools) for Igbo, a major Nigerian language. He was supervised by Dr. Mark Hepple who was interested in low resource language research and Igbo happened to be a good example. Their work focused on the design of the Igbo tagset, building a tagged corpus and developing the Igbo tagger. A fair amount of publications from different aspects of the project have been published.

In 2014, I joined the project, which we later tagged, *IgboNLP*. My focus was on a more subtle but quite problematic pre-processing challenge - 'diacritic restoration'. We observed that due to the high diacritic content in Igbo language, words with missing diacritics default to a base-form which becomes ambiguous without context. These ambiguities could be in meaning only but could also transcend word classes. We adopted a corpus-based approach that did not require any human annotation in building our experimental pipeline. The pipeline allows up to create training instances from raw data, apply one of the restoration methods we proposed i.e. n-gram, machine learning and embedding models. We made a few conference presentations with this work.

I joined the CorCenCC project in October of 2018 as a Research Associate with the WP3 team working on the development of the Welsh Semantic tagger. I got a huge insight from Steve Neale's and Scot Piao's excellent works on the ***CyTag*** and ***CySemTagger*** which are both rule-based taggers for Welsh part-of-speech and semantic tagging. Having done a bit of work with embedding models, I became interested in such meaning abstractions and semantic relationships as captured by deep embedding models. But as Kevin Donnelly pointed out, they require large amounts of data to be trained which are often not available for languages like Welsh making rule-based approaches a clear choice.

However, we are pushing the boundaries and focusing on finding ways to leverage existing pre-trained models which are often built from unstructured data using semi-supervised methods. We are collaborating with the WP2 team to work on training an embedding-based deep neural multi-tagger that can perform both part-of-speech and semantic tagging simultaneously. The work is on-going, but we have made substantial progress. Indeed, a paper on that will be presented at an ACL2019 workshop. We have also done a poster and demonstrated the tool at the CL2019 conference. We are motivated and very excited about our achievement so far and look forward to more contributions and achievements on the CorCenCC 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 email us at: 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