CorCenCC Newsletter

Issue 16, March 2018

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

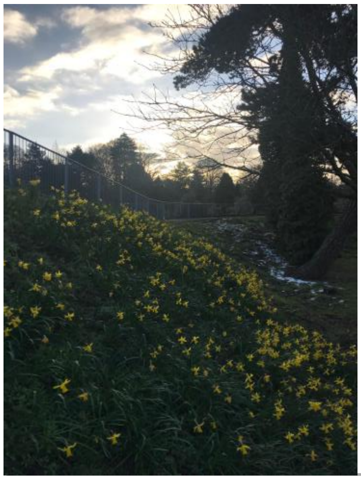
CyTag
Rule-based tagging for Welsh

Contents

Greetings from the PI

Welcome to springtime (well, let's hope we've seen the back of the snow now…)! March is finally here: the daffs are out, the lambs are gambolling and the team are still hard at work collecting data and building the foundations for the National Corpus of Contemporary Welsh!

This month marks the two-year anniversary of the start of the project which means we are now in the final furlong of the work. Over the next 18 months you will start to see a lot more activity from the team as some of the fruits of our labour start to get released. Numerous roadshows and demos of the data and corpus enquiry tools are also afoot - check out our website, social media feeds and future editions of the newsletter for more information!



The current newsletter brings you up-to-date with recent news from the project. This includes details about studentship successes; the launch of our part-of-speech tagger, *CyTag*, and some updates from Work Package 3 (WP3). As ever, you will get a chance to meet another member of the team (Tom Cobb) and, sadly, we also have to say goodbye to one…

Happy reading! Dr Dawn Knight

Farewell Dr Jeremy Evas

March sees changes on the CorCenCC project team. Dr Jeremy Evas is to be replaced by his friend and colleague at the School of Welsh, Dr Jonathan Morris. Jonathan and Jeremy recently collaborated on research into conditions influencing Welsh Language transmission and use in families (Welsh Government 2017). Jeremy is currently seconded out of the University as Head of Welsh Language Promotion at the Welsh Government. Jeremy said *"I'm sorry to be leaving such an interesting project and wish it all success with data collection and analysis. 'Parse' on my best wishes to the team"*

+ News and events

Welcome to the team: Dr Jonathan Morris

Stepping in to Jeremy's shoes on the project is Dr Jonathan Morris, who became a Co-Investigator (CI) on the team from 1st March 2018. Dr Jonathan Morris is the *Coleg Cymraeg Cenedlaethol* Lecturer in Linguistics and Applied Linguistics at the School of Welsh, Cardiff University. Jonathan is assisting with the coordination of data collection so will be contributing primarily to Work Package 1 (WP1). We asked Jonathan to provide a few words on joining the team and this is what he had to say:

Su' ma'i! I am a Coleg Cymraeg Cenedlaethol lecturer in linguistics and have worked at the School of Welsh, Cardiff University, for five years. I come from Wrexham originally but was brought up not far from the border in Cheshire. I liked languages at school and moved to Manchester to do a degree in French and German. As part of the degree, I spent time in Switzerland and started to become really interested in bilingualism. I completed a PhD on phonetic and phonological variation in the speech of Welsh-English bilinguals in 2013 and worked part time as a Welsh for Adults tutor.

At the School of Welsh, I teach on modules such as 'Language Acquisition' and 'Sociolinguistics' and do research on the social factors (such as area and linguistic background) which influence how Welsh speakers and learners speak in both of their languages. I also have an interest in the sociology of language and was part of a project on the transmission of Welsh in the home. Being a part of the CorCenCC project is really exciting – the project provides a unique opportunity for people of all backgrounds of from every corner of Wales to contribute to the future of the language.

Cardiff Undergraduate Research Opportunities Programme (CUROP) successes!

We are pleased to announce that we have received funding from the Cardiff Undergraduate Research Opportunities Programme (CUROP), for two student placements over the summer months. CUROP enables undergraduate students from Cardiff University to gain paid experience of academic research by applying to become involved in supervised projects over the summer months. The first of these projects is entitled Corpws Cenedlaethol Cymraeg Cyfoes: National Corpus of Contemporary Welsh – semantic tagging and data annotation. The CUROP researcher on this project will contribute to WP3 and work on the construction of the gold standard corpus and the study of related semantic issues.

The second project is entitled Corpws Cenedlaethol Cymraeg Cyfoes: National Corpus of Contemporary Welsh – a focus on spoken data. The CUROP researcher employed on this project will be supervised by the CorCenCC project lead, Dr Dawn Knight, in conjunction with Dr Lowri Williams, a Research Assistant (RA) on the project who is part of the team responsible for compiling the 10 million-word corpus. As you all know, we aim to collect data from contributors of different ages, geographical locations, genders and occupations for CorCenCC and the role of the second CUROP researcher on this project will be to contribute towards this aim. The successful applicant will assist with the following tasks: (1) Data collection, (2) Transcription and (3) Providing reflection and refinement on (1) and (2).

The CUROP researchers on both projects will also help to promote the CorCenCC project to the wider population, to encourage further contributors to the corpus and to generate interest in the potential future uses of the resource. We will be recruiting students to these CUROP projects the next couple of weeks. In the meantime, if you would like to learn more about the scheme, please see here: http://bit.ly/2pqN61A

Swansea SPIN (Swansea Paid Internship Network) placement: Croeso i Sioned!

SPIN placements are part of Swansea University's internship scheme, run by SEA (Swansea Employability Academy). Placements last for four weeks (or equivalent hours), are paid, and can be worked part time during term time or full time during vacation/non-teaching weeks. We were delighted that our application to SPIN for a CorCenCC placement was successful, and enabled us to advertise for a student to contribute to the project, concentrating on working with the Swansea team on (i) data collection and (ii) end-user interviews.

We were pleased to receive several strong applications, and can now announce that Sioned Johnson-Dowdeswell has been appointed and will be working with Steve, Tess, Jenny and Mair at Swansea University over the next few months.

Sioned will be working with members of the CorCenCC team to record Welsh speakers as they use the language in everyday situations across Wales (in particular in her home area of Carmarthen). She will also be helping us with transcription work and wider engagement efforts with the general public.

Work Package 4 (the pedagogic toolkit) will also benefit from Sioned's input. We anticipate that she will help us analyse data from interviews with teachers and tutors, collecting views that will help shape the development of pedagogical materials. This placement will add to collaboration between the Departments of Applied Linguistics and Welsh at Swansea and further strengthen links within WP1 and WP4 across all the institutions.

For more information on the SPIN programme, please visit: https://tinyurl.com/yaotju43

Welcome to the (ever-growing) CorCenCC team Sioned! Here's a word or two about who Sioned is:

"My name is Sioned. I am a first-year student at Swansea University studying Welsh. I am from Carmarthen and I continue to live there whilst studying at University. I am very fortunate to have this opportunity to work on the CorCencc project. Although I love the Welsh language and am hugely interested in its development, I also love animals, I have two dogs and five cats!!! The kitten in the photo is named Eddie after Ed Sheeran, for obvious reasons!! I also love cooking and photography in my spare time"

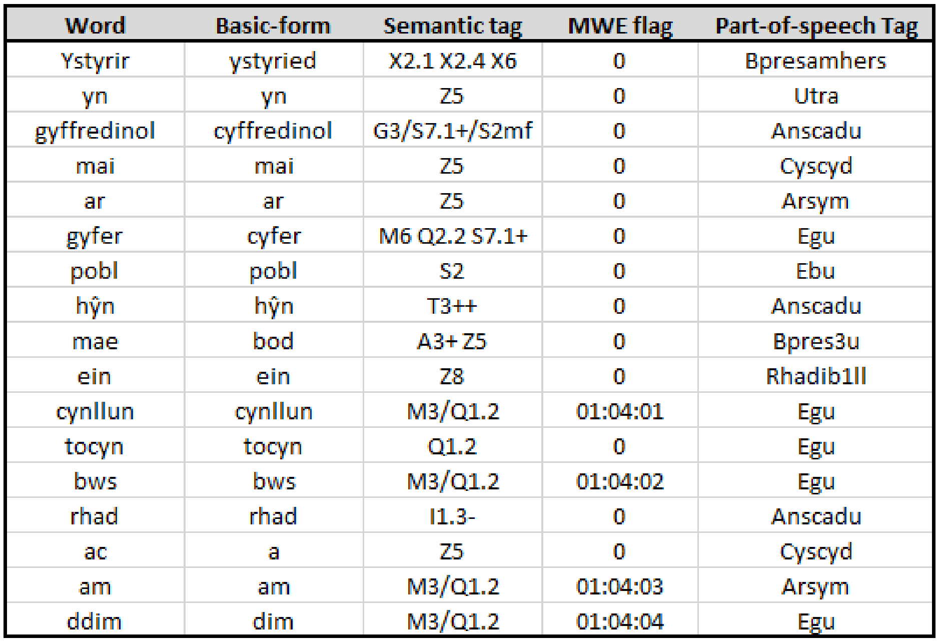
+ General updates

CySemTagger update

By Scott Piao and Paul Rayson

The development of Welsh Semantic tagger (named CySemTagger now) is in full swing again in WP3, with the recent rapid progress in creating a high-quality test dataset that will facilitate the development and evaluation of new methods. In this regard, RAs Lowri Williams and Jennifer Needs and other team members have done an excellent job in accurately annotating every word in the Gold Corpus with correct semantic category/ies from the USAS semantic scheme. Such a resource is critical for semantic tool development by allowing us to train and automatically annotated results against human experts' knowledge and judgement. In the following stage, various context-aware algorithms will be tested based on this test data to identify and integrate optimal methods for the CySemTagger to achieve a higher performance and accuracy.

The table to the right shows a semantically annotated sample of Welsh text. The third and fourth columns contain annotations provided by the CySemTagger, which respectively denote USAS semantic word is part of a multi-word expressions (including idioms and word groups that carry a single semantic concept). As an automatic tool, CySemTagger makes mistakes, and this is exactly where we need further improvement based on high-quality test data.



Word	Basic-form	Semantic tag	MWE flag	Part-of-speech Tag
Ystyrir	ystyried	X2.1 X2.4 X6	0	Bpresamhers
yn	yn	Z5	0	Utra
gyffredinol	cyffredinol	G3/S7.1+/S2mf	0	Anscadu
mai	mai	Z5	0	Cyscyd
ar	ar	Z5	0	Arsym
gyfer	cyfer	M6 Q2.2 S7.1+	0	Egu
pobl	pobl	S2	0	Ebu
hŷn	hŷn	T3++	0	Anscadu
mae	bod	A3+ Z5	0	Bpres3u
ein	ein	Z8	0	Rhadib1ll
cynllun	cynllun	M3/Q1.2	01:04:01	Egu
tocyn	tocyn	Q1.2	0	Egu
bws	bws	M3/Q1.2	01:04:02	Egu
rhad	rhad	I1.3-	0	Anscadu
ac	a	Z5	0	Cyscyd
am	am	M3/Q1.2	01:04:03	Arsym
ddim	dim	M3/Q1.2	01:04:04	Egu

CyTag: the launch

By Steven Neale

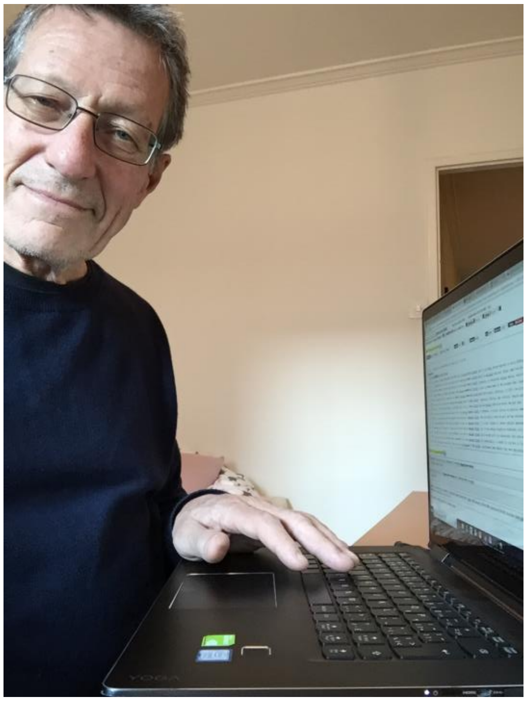
This month we're launching the website for CyTag, our open source pipeline of tagging tools for Welsh. CyTag has been developed as the major deliverable for WP2, whose aims were to create the part-of-speech (POS) tagger and bespoke tagset that will be used to tag CorCenCC. The URL for the website is http://cytag.corcencc.org, where you can find out more about the tagger and try out our online demo!

CyTag currently contains a text segmenter, a sentence splitter, a tokeniser, and the POS tagger itself. It's rule-based, and works by leveraging lexical information form the GPL dictionary Eurfa to find all the possible POS tags for a given token, before applying sequences of rules to the tokens that decide between those options based on the POS tags and/or the lexical features of the other tokens around it. We've been testing out CyTagusing a small collection of manually-annotated example sentences, and we're very happy with its accuracy - currently over 95%!

In more exciting news, we also had a paper describing CyTag accepted for publication at the upcoming Language Resources and Evaluation (LREC) 2018 conference in Miyazaki, Japan in May. The paper - authored by Steven Neale, Kevin Donnelly, Gareth Watkins and Dawn Knight - presents a thorough technical overview of CyTag and an in-depth evaluation of its accuracy. We're looking forward to presenting it to the wider community, and to getting some useful feedback to take forward!

+ Meet the team: Tom Cobb, University of Quebec at Montreal, Project Consultant

It is ironic that I am involved in this Welsh National Corpus project because although Canadian my whole career in education began in Wales – University College of North Wales, in Bangor, where I did my PGCE in English teaching in about 1978. This course involved going out on several teaching internships in schools, and the administrators of my college were forever looking for schools where a trainee without the slightest knowledge of Welsh language (me!) would be acceptable. These were typically within a few miles of the English border, usually in Clwyd. Over that period, however, I did learn a little Welsh, lived in a partially Welsh household, and frequented a pretty much Welsh-only pub.



However, it was only shortly after training to become a high school English teacher that I returned to Canada and became a sessional university lecturer in English composition, a job for which I was, of course, entirely untrained. I did however see that writing instruction had a large and unoccupied space for computer assisted learning (CALL), for example to practice various quasi-mechanical operations like applying punctuation and identifying sentence fragments. But I was not a software programmer at the time, so my speculations about a role for computing remained… theoretical.

The main other thing I learned in this Canadian university work was that it didn't pay very well, so I was shortly off to countries like Saudi Arabia and Oman where the salaries were better (and the conditions worse, but that's another story). About this time, I bought my first computer, and in the spacious hot Saudi afternoons when no one worked I started to learn computer programming. I did some simple things for learners, carrying a 'luggable' (not portable) suitcase-sized Osborne computer across the desert between college and parking so random students could experience the joys of CALL. Toward the end of this experience the Macintosh SE30 came out with Hypercard loaded and I was soon creating learning games and lots of other software and getting an occasional student to play with these a bit so I could identify the weak spots in the interfaces.

A bit later in a bid to remain in The Gulf but depart Saudi Arabia, I moved laterally to the pleasant desert Sultanate of Oman, for a similar job but with a Macintosh Lab to set up and run in a new college. This was perfect timing for my nascent coding skills, and in addition the Language Centre at Sultan Qaboos University was a hotbed of corpus and concordancing work (John Sinclair had even paid a visit, though before my time). It wasn't long before I had begun integrating corpus concepts into my games and drills, a project that led to a PhD in educational technology and many happy years puzzling out ways to engage learners in 'big data' as they work out solutions to the questions they have about language.

This corpus-and-CALL work led to a real job back in Canada and now has led to a role on the tutorial side of the Welsh Corpus project. Once you have a corpus, why not use it as a learning tool? Among other things, of course. For a language career that started in Wales, the symmetry could not be better!

+ 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@cardiff.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 and Jonathan Morris; **RAs** - Steven Neale, Jennifer Needs, Mair Rees, Scott Piao 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** - 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