CorCenCC Newsletter

Issue 1: April 2016

CorCenCC
Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Greetings from the PI

The CorCenCC project is now officially up and running! It all began on 1st March 2016 and will be running for the next 3.5 years (and is being funded by both ESRC and AHRC). The last five weeks have been very busy indeed. Members of the CorCenCC Project Team (CPT) have met and been focusing on one of the most important stages of a project of this nature: planning.

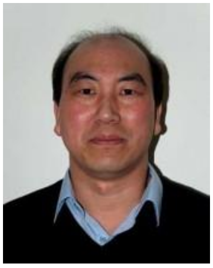
Aims and objectives from the bid have been unpacked, researchers have been trained, tasks and roles have been mapped out, and a clearer sense of what needs to be done, by whom and by when has been agreed. We have also had the opportunity to invite some of the consultants and advisory board members for the project to join discussions in person (Laurence Anthony and Kevin Donnelly) or via Skype (Tom Cobb) and hope to speak with other members of the wider CorCenCC project team in the near future too.

Although it is early days, I wanted to take this opportunity to update you on some of the latest news and developments of the project: introducing new members of the project team (RAs), welcoming the ambassadors of the project, updating you on the presence of CorCenCC online, and detailing early plans for conference presentations.

Happy reading, Dawn Knight

Introducing...new members of the CorCenCC team

We are proud to announce that we have employed 5 fantastic researchers to work on the CorCenCC project:



Dr. Scott Piao has rich experience in corpus tool development and natural language processing. He has worked on seven projects funded by EPSRC, AHRC, EU and JISC, covering research topics of corpus construction and annotation, text mining, and application of corpus and natural language processing techniques in social computing. In particular, he has been involved in the development of corpus semantic annotation system for multiple languages for many years. Scott is contributing to the development of the CorCenCC semantic tagger and crowdsourcing utilities.

Dr Steven Neale joined Cardiff University in March 2016 as a Research Associate on the CorCenCC project. Before coming to Cardiff he was a Postdoctoral Researcher with the NLX – Natural Language and Speech group at the University of Lisbon in Portugal, where he had been working since 2014 after completing his PhD in Computing at the University of Tasmania in Australia. Before deciding to pursue an academic career, Steven spent his first years after university working in various film, TV and video production jobs. Steven is working with Irena and Dawn to build and operationalize the part-of-speech tagger, crowdsourcing applications, pedagogic toolkit and corpus infrastructure.

Dr Gareth Watkins graduated with a Degree in Welsh from Swansea University in 1999. He returned to Swansea University in 2007 to follow an MA in Translation with Language Technology which he completed in October 2008. After graduating, Gareth delivered a Translation and Technology module to 2nd year undergraduates on the BSc Translation course at Aston University, before returning to Swansea University again, this time to study towards a PhD. Shortly after obtaining the PhD he delivered Undergraduate and Postgraduate level modules on Language Technology using on-line distance learning methods for Swansea University as part of the Training in Languages and Translation (TILT) Project. Gareth is contributing to the collection of the Welsh language data for CorCenCC, developing a part-of-speech tagset and providing language-specific expertise.

Dr Jennifer Needs has recently completed a PhD in the School of Welsh at Cardiff University, in which she looked at the principles of online language learning materials development. She used these principles to inform the development of her own e-learning materials for adults learning Welsh, working in partnership with the Nant Gwrtheyrn Welsh Language and Heritage Centre to create bespoke online materials for their learners. Prior to her PhD, Jennifer worked as a Research Assistant at Cardiff University in the field of Welsh for Adults, and completed a BA in Linguistics and Spanish at Leeds University and an MA in Endangered Language Documentation and Revitalisation at SOAS, London. Jennifer is contributing to the collection of the Welsh language data for CorCenCC, developing the pedagogic toolkit and providing language-specific expertise throughout.

Following a 15 year career as an art therapist working mainly with adults who have a learning disability, **Dr Mair Rees** returned to full-time education to study for a BA Welsh at Cardiff University in 2004. Subsequently, she was fortunate enough to win a scholarship which also enabled her to undertake a PhD in Welsh literature. Since graduating in 2012 Mair has worked as a creative editor with Gomer Press, Llandysul. She regularly contributes reviews and articles to Welsh journals and also has a small business making quirky Welsh-language gifts and cards. Mair is contributing to the collection of the Welsh language data and providing language-specific expertise throughout the project.

Project ambassadors

It's a pleasure to announce that **Nia Parry** (TV presenter, producer and researcher; Welsh tutor, *Welsh in a week* (S4C)); **Nigel Owens** (international rugby referee; TV presenter), **Cerys Matthews** (Musician author; radio and TV presenter) and **Damian Walford Davies** (Prof. of English Literature; poet Chair of Literature Wales) are the official ambassadors of the CorCenCC project.

CorCenCC online

We have set up a holding website for the project. This is a static website which outlines the basic aims of the project and provides details of team members and contact information: http://sites.cardiff.ac.uk/corcencc/

The main project website (that will host the corpus) will be launched later this year: www.corcencc.org

You can also 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

CorCenCC @ conferences

Despite being very early days, I am pleased to announce that the following papers have been accepted for presentations at conferences in 2016:

- **Fitzpatrick, T., Knight, D.** and **Morris, S.** (2016). Creating pedagogical wordlists: a comparison of thematic and corpus approaches. Paper to be presented at the *Pacific Second Language Research Forum (PacSLRF2016)*, September 2016, Tokyo, Japan.
- **Knight, D., Fitzpatrick, T.** and **Morris, S.** (2016). CorCenCC - Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh). *WISERD (Wales Institute of Social and Economic Research, Data and Methods)*, July 2016, Swansea University.
- **Knight, D., Neale, S., Spasic, I., Morris, S.** and **Fitzpatrick, T.** (2016). Crowdsourcing corpus construction: contextualizing plans for CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes - The National Corpus of Contemporary Welsh). Paper to be presented at the *IVACS 2016 conference*, June 2016, Bath Spa University.
- **Needs, J., Rees, M., Watkins, G., Morris, S., Knight, D.** and **Fitzpatrick, T.** (2016). CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – The National Corpus of Contemporary Welsh): Challenges and applications in a minoritised language context. Paper to be presented at the *IVACS 2016 conference*, June 2016, Bath Spa University.
- **Piao, S., Rayson, P.,** Archer, D., Bianchi, F., Dayrell, C. El-Haj, M., Jiménez, R-M., **Knight, D.,** Michal Křen, M., Löfberg, L., Nawab, R., Shafi, J., The, P-L. and Mudraya, O. (2016). Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. Paper delivered at the *LREC (Language Resources Evaluation) 2016 Conference,* May 2016, Slovenia.

If you have any comments or questions about the content of this newsletter please contact Dr Dawn Knight: KnightD5@cardiff.ac.uk

Cylchlythyr CorCenCC

Rhifyn 1: Ebrill 2016

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Cyfarchion gan y Prif Ymchwilydd

Mae prosiect CorCenCC bellach ar waith! Dechreuodd popeth ar 1 Mawrth 2016, a chaiff y prosiect ei gynnal am y 3.5 blynedd nesaf (a'i ariannu gan yr ESRC ac AHRC). Mae'r pump wythnos diwethaf wedi bod yn brysur iawn. Mae aelodau o Dîm Prosiect CorCenCC (CPT) wedi cwrdd, ac wedi bod yn canolbwyntio ar un o gamau pwysicaf prosiect o'r fath: cynllunio.

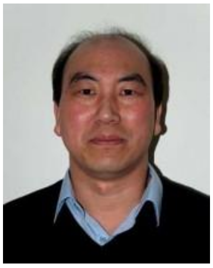
Mae nodau ac amcanion y cais wedi'u hystyried, mae ymchwilwyr wedi'u hyfforddi, mae tasgau a rolau wedi'u hamlinellu, ac mae gan bawb well syniad o beth sydd angen ei wneud, gan bwy ac erbyn pryd. Cawsom hefyd gyfle i wahodd rhai o'r ymgynghorwyr ac aelodau'r bwrdd cynghori ar gyfer y prosiect i ymuno â thrafodaethau'n bersonol (Laurence Anthony a Kevin Donnelly) neu drwy Skype (Tom Cobb), ac rydym yn gobeithio siarad ag aelodau eraill o dîm ehangach prosiect CorCenCC yn y dyfodol agos hefyd.

Er mai megis dechrau yr ydym, roeddwn am fanteisio ar y cyfle hwn i roi'r wybodaeth ddiweddaraf i chi am newyddion a datblygiadau'r prosiect: cyflwyno aelodau newydd o dîm y prosiect (Cynorthwywyr Ymchwil), croesawu llysgenhadon y prosiect, rhoi'r wybodaeth ddiweddaraf i chi am bresenoldeb CorCenCC ar-lein, a rhoi manylion cynlluniau cynnar ar gyfer cyflwyniadau mewn cynadleddau.

Mwynhewch, Dawn Knight

Cyflwyno aelodau newydd o dîm CorCenCC

Rydym yn falch o gyhoeddi ein bod wedi cyflogi 5 o ymchwilwyr gwych i weithio ar brosiect CorCenCC:



Mae gan **Dr Scott Piao** brofiad helaeth o ddatblygu offer corpws a phrosesu iaith naturiol. Mae wedi gweithio ar saith prosiect a ariennir gan EPSRC, AHRC, yr UE a JISC. Mae'r pynciau ymchwil o dan sylw wedi cynnwys creu ac anodi corpws, chwilio am destun, a defnyddio technegau prosesu iaith naturiol a chorpws mewn cyfrifiadura cymdeithasol. Yn benodol, mae wedi bod yn gysylltiedig â datblygu system anodi semantig ar gyfer corpora ar gyfer sawl iaith dros flynyddoedd lawer. Mae Scott yn cyfrannu at ddatblygu tagiwr semantig CorCenCC a'r adnoddau torfoli.

Ymunodd **Dr Steven Neale** â Phrifysgol Caerdydd ym mis Mawrth 2016 fel Cydymaith Ymchwil ar y prosiect CorCenCC. Cyn dod i Gaerdydd roedd yn Ymchwilydd Ôl-ddoethuriaeth gyda NLX – grŵp Iaith a Lleferydd Naturiol ym Mhrifysgol Lisbon, Portiwgal, lle bu'n gweithio ers 2014 wedi iddo gwblhau ei PhD mewn Cyfrifiadura ym Mhrifysgol Tasmania yn Awstralia. Cyn penderfynu mynd ar drywydd gyrfa academaidd, treuliodd Steven ei flynyddoedd cyntaf wedi'r Brifysgol yn gweithio mewn amrywiol swyddi cynhyrchu ffilmiau, teledu a fideo. Mae Steven yn gweithio gydag Irena a Dawn i adeiladu a gweithredu'r tagiwr rhannau ymadrodd, yr apiau torfoli, y pecyn cymorth pedagogaidd ac isadeiledd y corpws.

Graddiodd **Dr Gareth Watkins** gyda gradd yn y Gymraeg o Brifysgol Abertawe yn 1999. Dychwelodd i Brifysgol Abertawe yn 2007 i ddilyn MA mewn Cyfieithu gyda Thechnoleg Iaith ac fe'i cwblhaodd ym mis Hydref 2008. Ar ôl graddio, cyflwynodd Gareth fodiwl Cyfieithu a Thechnoleg i israddedigion 2il flwyddyn ar y cwrs BSc Cyfieithu ym Mhrifysgol Aston, cyn dychwelyd i Brifysgol Abertawe unwaith eto, y tro hwn i astudio tuag at PhD. Yn fuan ar ôl cael y PhD cyflwynodd fodiwlau lefel Israddedig ac Ôl-raddedig ar Dechnoleg Iaith gan ddefnyddio dulliau dysgu o bell ar-lein ar gyfer Prifysgol Abertawe fel rhan o'r Prosiect Hyfforddiant mewn Ieithoedd a Chyfieithu (TILT). Mae Gareth yn cyfrannu at gasglu'r data Cymraeg ar gyfer CorCenCC, datblygu set tagiau rhannau ymadrodd a darparu arbenigedd iaith penodol.

Mae **Dr Jennifer Needs** wedi cwblhau PhD yn Ysgol y Gymraeg ym Mhrifysgol Caerdydd yn ddiweddar. Yn ei gwaith PhD edrychodd ar egwyddorion datblygiad deunyddiau dysgu iaith ar-lein. Defnyddiodd dair egwyddor fel sail i ddatblygiad ei deunyddiau e-ddysgu ei hun i oedolion sydd yn dysgu Cymraeg, gan weithio mewn partneriaeth â Chanolfan Iaith Gymraeg a Threftadaeth Nant Gwrtheyrn i greu deunyddiau ar-lein unigryw i ddysgwyr. Cyn ei PhD, roedd Jennifer yn gweithio fel Cynorthwyydd Ymchwil ym Mhrifysgol Caerdydd ym maes Cymraeg i Oedolion, a chyflawnodd radd BA mewn Ieithyddiaeth a Sbaeneg ym Mhrifysgol Leeds ac MA mewn Dogfennu ac Adfywio Ieithoedd mewn Perygl yn SOAS, Llundain. Jennifer sy'n cyfrannu at gasglu'r data Cymraeg ar gyfer CorCenCC, datblygu'r pecyn cymorth pedagogaidd a darparu arbenigedd iaith penodol drwy gydol y prosiect.

Yn dilyn gyrfa 15 mlynedd fel therapydd celf yn gweithio'n bennaf gydag oedolion ag anabledd dysgu, dychwelodd **Dr Mair Rees** i addysg llawn amser i astudio am radd BA mewn Cymraeg ym Mhrifysgol Caerdydd yn 2004. Wedyn, bu'n ddigon ffodus i ennill ysgoloriaeth a wnaeth hefyd ei galluogi i wneud PhD mewn Llenyddiaeth Gymraeg. Ers iddi raddio yn 2012 mae Mair wedi gweithio fel golygydd creadigol gyda Gwasg Gomer, Llandysul. Mae'n cyfrannu'n rheolaidd at adolygiadau ac erthyglau i gylchgronau Cymraeg ac mae ganddi hefyd fusnes bach yn gwneud cardiau ac anrhegion Cymraeg ychydig yn wahanol. Mae Mair yn cyfrannu at gasglu'r data Cymraeg a darparu arbenigedd iaith drwy gydol y prosiect.

Llysgenhadon y prosiect

Mae'n bleser cael cyhoeddi mai **Nia Parry** (cyflwynydd teledu, cynhyrchydd ac ymchwilydd; tiwtor Cymraeg, *Welsh in a week* (S4C)); **Nigel Owens** (dyfarnwr rygbi rhyngwladol; cyflwynydd teledu), **Cerys Matthews** (cerddor; awdur; cyflwynydd radio a theledu) a **Damien Walford Davies** (Athro Llenyddiaeth Saesneg; bardd; Cadeirydd Llenyddiaeth Cymru) yw llysgenhadon swyddogol prosiect CorCenCC.

CorCenCC ar-lein

Rydym wedi sefydlu gwefan dros dro ar gyfer y prosiect. Gwefan statig yw hon sy'n amlinellu nodau sylfaenol y prosiect ac yn rhoi manylion am aelodau'r tîm a manylion cyswllt: http://sites.caerdydd.ac.uk/corcencc/

Bydd prif wefan y prosiect (a fydd yn gartref i'r corpws) yn cael ei lansio yn ddiweddarach eleni: www.corcencc.org

Mae'r wybodaeth ddiweddaraf am ddatblygiadau'r prosiect hefyd ar gael drwy Facebook www.facebook.com/CorCenCC/; Twitter https://twitter.com/corcencc (gallwch ein trydar @CorCenCC). Gallwch hefyd gysylltu â ni drwy anfon neges i gyfeiriad ebost y prosiect: corcencc@caerdydd.ac.uk

CorCenCC mewn cynadleddau

Er mai megis dechrau yr ydym, rwy'n falch o gyhoeddi bod y papurau canlynol wedi'u derbyn ar gyfer cyflwyniadau mewn cynadleddau yn 2016:

- **Fitzpatrick, T., Knight, D.** a **Morris, S.** (2016). Creating pedagogical wordlists: a comparison of thematic and corpus approaches. Papur i'w gyflwyno yn *Pacific Second Language Research Forum* (PacSLRF2016), Medi 2016, Tokyo, Japan.
- **Knight, D., Fitzpatrick, T.** a **Morris, S.** (2016). CorCenCC - Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh). *WISERD (Sefydliad Ymchwil Gymdeithasol ac Economaidd, Data a Dulliau Cymru)*, Gorffennaf 2016, Prifysgol Abertawe.
- **Knight, D., Neale, S., Spasic, I., Morris, S.** a **Fitzpatrick, T.** (2016). Crowdsourcing corpus construction: contextualizing plans for CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes - The National Corpus of Contemporary Welsh). Papur i'w gyflwyno yng nghynhadledd *IVACS 2016*, Mehefin 2016, Prifysgol Spa Caerfaddon.
- **Needs, J., Rees, M., Watkins, G., Morris, S., Knight, D.** a **Fitzpatrick, T.** (2016). CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – The National Corpus of Contemporary Welsh): Challenges and applications in a minoritised language context. Papur i'w gyflwyno yng nghynhadledd *IVACS 2016*, Mehefin 2016, Prifysgol Spa Caerfaddon.
- **Piao, S., Rayson, P.,** Archer, D., Bianchi, F., Dayrell, C. El-Haj, M., Jiménez, R-M., **Knight, D.,** Michal Křen, M., Löfberg, L., Nawab, R., Shafi, J., The, P-L. a Mudraya, O. (2016). Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. Papur i'w gyflwyno yng nghynhadledd *LREC (Language Resources Evaluation) 2016*, Mai 2016, Slofenia.

Os oes gennych unrhyw sylwadau neu gwestiynau am gynnwys y cylchlythyr hwn, cysylltwch â Dr Dawn Knight: KnightD5@caerdydd.ac.uk