CorCenCC

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Cynnwys

T2: Newyddion a digwyddiadau

T4: Diweddariadau cyffredinol

T5: Cwrdd â'r tîm

T6: Cysylltu â ni

Cyfarchion gan y Prif Ymchwilydd

Blwyddyn Newydd Dda i'n darllenwyr i gyd! Gobeithio i chi fwynhau gwyliau'r Nadolig (er eu bod yn teimlo'n bell i ffwrdd erbyn hyn!) a gobeithio y bydd dechrau 2018 yn gynhyrchiol i bawb. Mae pethau wedi cychwyn yn dda yma ym mhencadlys CorCenCC ac ar draws ein sefydliadau partner: mae cynlluniau ar gyfer blwyddyn brysur ar y gweill. Dros 20 mis nesaf y prosiect bydd ffrwyth ein llafur yn dod yn fwy amlwg i chi i gyd gyda chynlluniau ar droed i ryddhau fersiynau cynnar o'r corpws a'r offer sy'n gysylltiedig ag ef. Rydym yn cynllunio nifer o sioeau teithio a digwyddiadau i'n helpu i broffilio CorCenCC a'i ddangos ar waith – felly mynnwch gip ar rifynnau nesaf y cylchlythyr, y dudalen Facebook, negeseuon Trydar a'r wefan am fwy o fanylion a galwch heibio i gadw cwmni â ni! Ar ben hynny, byddwn yn canolbwyntio ar ledaenu gwybodaeth am ein gwaith mewn cynadleddau a chyfarfodydd gan obeithio mynychu'r Eisteddfod a Thafwyl eto, felly bydd cyfle i ni gwrdd â chi yn un o'r digwyddiadau hynny hefyd! Mae'r rhifyn, eitha cryno, yma o'r cylchlythyr yn rhoi'r newyddion diweddaraf i chi o'r prosiect. Mae hyn yn cynnwys manylion am deithiau i Ffrainc ac i Siapan yn y dyfodol; ambell ddiweddariad ar ble rydyn ni arni o ran trawsgrifio a chrafu'r we (er mwyn casglu data iaith electronig) a manylion am Gyfarfod y Prosiect Cyfan sydd ar y gweill. Fel ym mhob rhifyn, fe gewch chi gyfle i gwrdd ag aelod arall o'r tîm (Paul Rayson) a, gwaetha'r modd, mae'n rhaid i ni ffarwelio hefyd ag un o'n cyd-ymchwilwyr ar y prosiect, Mark Stonelake, wrth iddo ymddeol.

Mwynhewch y cylchlythyr! Dr Dawn Knight

CorCenCC yn Ffrainc

Mae arweinydd y prosiect, Dawn Knight, wedi cael ei gwahodd i roi papur yng nghyfarfod llawn gweithdy'r Association Française de Linguistique Cognitive (Cymdeithas Ieithyddiaeth Wybyddol Ffrainc – AFLiCo) ar Gorpora a Chynrychioldeb. Fe'i cynhelir yn Nanterre ar 3-4 Mai 2018. Nod y gweithdy yw ymchwilio i'r hyn mae cynrychioldeb yn ei feddwl wrth ddylunio corpws; ydy hyn yn rhywbeth y mae modd ei gyflawni mewn gwirionedd; sut gallwn ni fynd i'r afael â thuedd mewn corpora ac ai'r un peth yw cynrychioldeb a chydbwysedd. Enw papur Dawn fydd 'Representativeness in CorCenCC: corpus design in minoritised languages'

Newyddion a digwyddiadau

Cyfarfod Tîm Cyfan y Prosiect @ Prifysgol Caerdydd

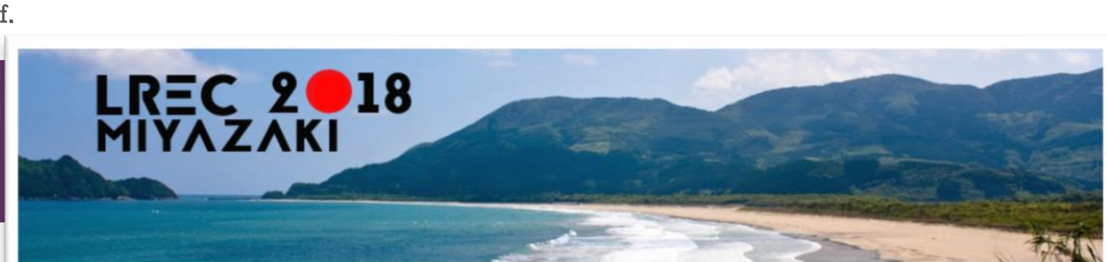
Cynhelir ail gyfarfod Tîm Prosiect Cyfan CorCenCC ym Mhrifysgol Caerdydd ar 26ain Chwefror 2018. Gwahoddir pob aelod o dîm prosiect CorCenCC, sy'n cynnwys 30 o aelodau, i'r cyfarfod dwyieithog hwn, fydd yn rhoi cyfle i'r rheini sy'n bresennol ystyried sut mae holl becynnau gwaith y prosiect yn datblygu, a chyfrannu at gynllunio yn y dyfodol. Byddwn yn trafod amrywiaeth o bynciau gan gynnwys sut y gallwn, o bosibl, gynhyrchu a chofnodi 'effaith' CorCenCC; cynlluniau ar gyfer dyfodol y corpws; beth yw'r ffyrdd gorau o ymgysylltu â defnyddwyr yn y dyfodol, a pha estyniadau a phrosiectau lloeren o CorCenCC fydd yn bosibl yn y dyfodol. Edrychwn ymlaen at weld pob aelod o'r tîm yng Nghaerdydd!

LREC 2018

Mae'r papurau canlynol gan aelodau tîm CorCenCC wedi eu derbyn i'w cyflwyno yn yr 11eg Cynhadledd Adnoddau Ieithyddol a Gwerthuso (LREC), fydd yn cael ei chynnal rhwng 7fed a 12fed Mai yn Miyazaki, Japan:

- Piao, S., Rayson, P., Knight, D. a Watkins, G. 2018 Towards a Welsh Semantic Annotation System. *Gweithgareddau'r Gynhadledd Adnoddau Ieithyddol a Gwerthuso (LREC) 2018*, Mai 2018, Miyazaki, Japan.
- Neale, S., Donnelly, K., Watkins, G. a Knight, D. (2018). Leveraging Lexical Resources and Constraint Grammar for Rule-Based Part-of-Speech Tagging in Welsh. Gweithgareddau'r Gynhadledd Adnoddau Ieithyddol a Gwerthuso (LREC) 2018, Mai 2018, Miyazaki, Japan.

Mae LREC yn gynhadledd hynod o gystadleuol a mawreddog sy'n canolbwyntio ar ddulliau cyfrifiadurol a chorpws arloesol, a datblygu adnoddau. Bydd yn cynnwys gweithdai, posteri a phapurau ar amrywiaeth o themâu, o ddadansoddi ymdeimladau ac adnabod emosiwn a chorpora ar gyfer dadansoddi iaith, i ddysgu iaith trwy gymorth cyfrifiadur. Mae crynodebau'n cael eu hadolygu gan gymheiriaid mewn modd 'dwbl-ddall' ac ychydig iawn sy'n cael eu derbyn. Gan mai cynhadledd am ieithyddiaeth gyfrifiadurol yw hon, mae pob un o'r cyflwyniadau hyn yn cynnwys crynodeb estynedig a fydd yn arwain at gyhoeddi yn ystod gweithgareddau'r gynhadledd. Bydd mynd i'r gynhadledd yn ein galluogi i gyflwyno rhai canfyddiadau allweddol sy'n deillio o fy ngwaith ar WP3 a WP2 CorCenCC ar gyfer y gymuned gyfrifiadurol/corpws ehangach, ac amlygu amcanion y prosiect yn fwy eang. Rydym hefyd yn gobeithio cael y cyfle i siarad ag ymgynghorydd y prosiect, Laurence Anthony, pan fyddwn yn Japan. Byddwn yn cyflwyno adroddiad am sut aiff pethau yn y gynhadledd hon, nes ymlaen yn ystod yr haf.



Mae'n drist iawn gorfod cyhoeddi y bydd Mark Stonelake, Cyd-Ymchwilydd ar brosiect CorCenCC, yn gadael y tîm pan fydd yn ymddeol o Brifysgol Abertawe. Fel awdur deunyddiau ac athro iaith profiadol, mae Mark wedi bod yn aelod gwerthfawr iawn o dîm y prosiect; byddwn yn gweld eisiau'r mewnwelediad a'r brwdfrydedd mae'n eu cyfrannu i gyfarfodydd ein tîm. Dymunwn ddweud 'Diolch yn Fawr' enfawr iddo am ei holl waith caled ac anfon pob dymuniad da iddo ar gyfer ei ymddeoliad. Rydym yn gwybod y bydd brwdfrydedd Marc tuag at amcanion ac ethos gwaith CorCenCC yn parhau, ac rydym yn gobeithio y bydd yn dal i ymwneud â'r prosiect wrth iddo ddatblygu – paid bod yn ddiarth Mark! Gofynnon ni i Mark ysgrifennu cwpwl o eiriau am ei ymddeoliad – dyma beth ddywedodd e:

Mark Stonelake yn ymddeol

Faint o'r gloch yw hi? Ychydig o flynyddoedd yn ôl, dechreuais gwrs gradd ym Mhrifysgol Abertawe. Wel, dyna sut mae'n teimlo, beth bynnag. Ond nôl yn 1986 oedd hynny! Ro'n i newydd roi'r gorau i'm swydd fel arborydd/garddwr i wneud BA Cymraeg. Bryd hynny, roedd myfyrwyr yn derbyn grant, nid benthyciad, a finnau'n fyfyriwr hŷn, cefais grant gwell. Ond oedden nhw'n ddyddiau da! Pe na bai grantiau ar gael bryd hynny, a ninnau'n magu plant bach, dw i ddim yn credu y byddwn wedi mynd amdani. Serch y grant gwell, roedd arian yn brin, felly pan ofynnwyd i fi ddysgu dosbarth nos, neidiais i at y cyfle. Doedd pobl ddim yn poeni am gymwysterau gymaint ers lawer dydd. Ers achub ar y cyfle cyntaf, rwy'n gweithio, heblaw am gyfnod byr fel tiwtor rhan amser a gŵr tŷ (os taw dyna yw'r term cywir am y peth erbyn hyn), mewn amryw swyddi llawn amser i Gymraeg i Oedolion.

Rwy wedi mwynhau pob agwedd ar y gwaith, o ddysgu a datblygu deunyddiau i hyfforddi a rheoli staff. Mae'r cyfle i arwain llawer o gyrsiau yng Nghymru ac America ac ymweliadau â'r Ffindir, Catalwnia a Gwlad y Basg wedi bod yn wych. Cefais gyfle i gwrdd â phobl o wahanol gefndiroedd a gwledydd ar y cyrsiau gwahanol, ac fel rhan o dîm CorCenCC. Allwn i byth wedi gwneud hyn i gyd heb gymryd y risg o fynd i'r brifysgol a newid fy ngyrfa nôl yn yr wythdegau.

Gobeithio mod i wedi gwneud cyfraniad gwerthfawr at Gymraeg i Oedolion dros y blynyddoedd a mod i wedi gwneud cyfraniad bach at brosiect CorCenCC hefyd. Fodd bynnag, daeth yr amser i roi heibio'r bwrdd gwyn a'r taenlenni. Byddaf yn ymddeol o'm swydd fel Rheolwr Datblygu Proffesiynol ac Ansawdd ddiwedd mis Chwefror. Cofiaf ein cyflawniadau a'n gwaith caled. Bydd gennyf gofion cynnes am fy myfyrwyr a'm cyd-weithwyr. Wel y rhan fwyaf helaeth ohonynt, beth bynnag. Ond y peth mwyaf cofiaf fydd y chwerthin a'r hwyl. Y rhan fwyaf o'r amser, dw i wedi cael cymaint o hwyl, fel nad oedd y gwaith yn teimlo fel gwaith o gwbl.

Gobeithio byddaf yn dysgu ambell i ddosbarth Cymraeg o hyd, fel tiwtor cysylltiol. Gobeithiaf hefyd gymryd rhan yn natblygiad y corpws mewn rhyw ffordd na'i gilydd. Felly, ry'ch chi'n debygol o'm gweld o gwmpas. Felly, Mark, gyda'ch gwybodaeth helaeth a'ch doethineb enfawr a enillwyd drwy waith llafurus ar y ffás ym myd addysg, oes gennych unrhyw gyngor i ni? Wel, gan i chi ofyn, credaf y dylem roi sylw manwl i eiriau doeth 'Mr Blwyddyn Newydd' ei hun - Guy Lombardo::

Enjoy yourself, it's later than you think
Enjoy yourself, while you're still in the pink
The years go by, as quickly as you wink
Enjoy yourself, enjoy yourself, it's later than you think

Rwyf wedi mwynhau gweithio gyda chi i gyd a dymunaf bob llwyddiant a hapusrwydd ar gyfer y dyfodol.

Mark Stonelake, e.m.stonelake@swansea.ac.uk

Casglu data a recriwtio cyfranogwyr

Rydym yn gobeithio dechrau'r flwyddyn newydd yn lledaenu'r gair am yr hyn sydd wedi bod yn digwydd yma ar CorCenCC. Bydd cyfryngau cymdeithasol yn chwarae rhan bwysig, ac rydym wedi bod yn meddwl yn galed am y ffordd orau o ddefnyddio ein sianeli Facebook a Twitter i wneud hyn.

Felly beth sydd wedi bod yn digwydd … a beth sy'n dod nesaf?

Rhannu - Y peth gwych am y prosiect yw gallwch gyfrannu gartre oddi ar eich ffôn symudol. Mae’r dyddiau lle'r oedd angen technoleg ffansi i recordio bellach wedi mynd. Popeth sydd ei angen arnoch yw iPhone a rhywle a rhywun i’w recordio (er ein bod ni’n dal i ddefnyddio’r rhain a gallan nhw fod o help mawr!). Y cyfan sydd ei angen nawr yw iPhone a rhywle (a rhywun) i recordio. Felly, byddwn yn annog pobl, trwy Facebook a Twitter, i lawrlwytho'r ap iPhone a chymryd rhan. Edrychwch am y hashnodau #WeWantYourWelsh #RhowchEichCymraegINi a helpwch ni drwy eu rhannu. Gorau po fwyaf!

Tîm sy’n tyfu - Rydym bob amser yn chwilio am drawsgrifwyr newydd, ac mae'r cyfryngau cymdeithasol wedi bod yn ffordd wych o’u recriwtio. Cyn y Nadolig ac yn ddiweddar, rydym wedi bod yn galw am drawsgrifwyr ar Facebook ac ar Twitter, ac mae’r ymateb wedi bod yn wych. Hyd yn hyn, mae 60 o bobl wedi dangos diddordeb mewn gweithio i ni, gyda cheisiadau newydd bob dydd. Rydym yn gyffrous iawn am yr hyn y gallwn ei gyflawni gyda’n gilydd eleni!

Yn dod yn fuan i siop goffi / siop trin gwallt / ysgol yn agos atoch chi … - Mae ein tudalennau Facebook a Twitter yn ffordd wych o ddarganfod ble bydd tîm CorCenCC yn recordio. Byddwn yn rhoi gwybod i chi os bydd aelod o’r tîm yn dod i ardal neu ddigwyddiad penodol a allai fod o ddiddordeb i chi. A oes rhywun y mae'n rhaid i ni sgwrsio â nhw? A oes yna ddigwyddiad y credwch y byddai'n wych ar gyfer y corpws? Gadewch i ni wybod! Dydyn ni ddim yn cnoi!

Felly cadwch lygad ar yr holl ddigwyddiadau eleni, yn enwedig ar ein cyfrifon cyfryngau cymdeithasol. Mae prosiect CorCenCC yn ymwneud â chydweithio ac arloesi, byddem wrth ein boddau i chi gymryd rhan. Dilynwch brosiect CorCenCC ar Facebook neu ar Twitter @corcencc.

+ Diweddariad cyffredinol: Casglu E-iaith

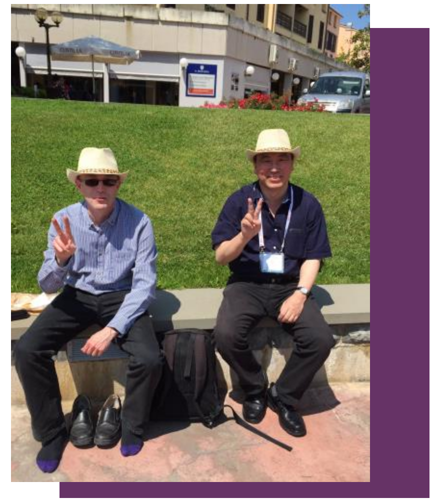
We-grafu ar waith

Gan Lowri Williams

Ers ein diweddariad diwethaf, un prif ffocws ar gyfer WP1 yw casglu data o wefannau a mathau eraill o destun electronig. Ein nod yw casglu 2 miliwn o eiriau a ddefnyddir mewn cyd-destunau o'r fath. Y dasg gyntaf oedd nodi gwefannau neu flogiau a gyflwynodd eu testun trwy gyfrwng y Gymraeg. Rydym wedi bod yn cysylltu â pherchnogion gwefannau neu flogwyr i ofyn am eu caniatâd i gynnwys eu deunydd yn ein corpws. Ar hyn o bryd mae gennym gyfanswm o 152 o wefannau a blogiau sydd wedi cytuno i gyfrannu eu cyd-destunau i CorCenCC. Mae nifer fawr o safleoedd yn golygu bod llawer o ddata i'w gasglu. O gofio bod fy nghefndir mewn cyfrifiadureg, roeddwn i'n gallu defnyddio fy sgiliau i weithredu ‘scrapiwr gwe’ i dynnu testunau yn awtomatig. Ysgrifennir y scrapiwr gwe mewn Python, iaith raglennu. Y gofyniad cyntaf yw nodi pa dudalennau o wefan sydd i'w defnyddio. Gan ddefnyddio nifer o lyfrgelloedd Python, gallwn nodi URL gwefan a dychwelyd yr holl dudalennau sy'n weddill ar gyfer y safle a roddir. Yna, caiff y URLs eu mewnbynnu i'r ‘scrapiwr’ sy'n dychwelyd ac yn glanhau'r HTML, yr iaith ar gyfer creu tudalennau gwe. Gallwn gasglu testun gwefan mewn llai na 5 munud. Cyfanswm y nifer o eiriau sydd wedi cael eu casglu hyd yn hyn yw 1.5 miliwn. Y dasg nesaf yw anonymeiddio'r data er mwyn iddo gael ei gynnwys yn y corpws terfynol.

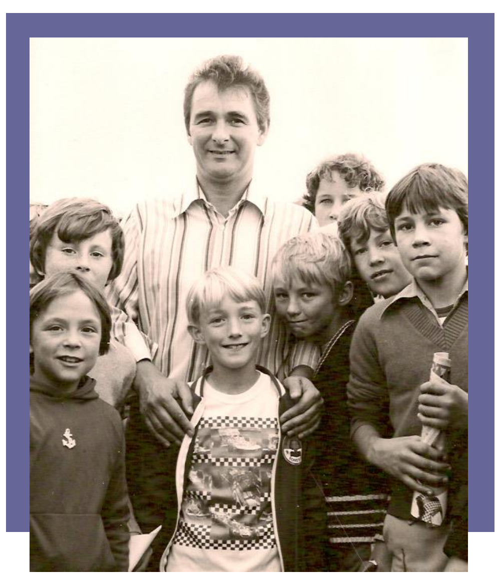
+ Cwrdd â'r tîm:

Dr Paul Rayson, Cyd-Ymchwilydd, Prifysgol Caerhirfryn



Felly, i ddechrau, un o ble ydw i? Cefais fy ngeni yn y chwedegau (cyn i Neil Armstrong gerdded ar y lleuad!). Roeddwn yn byw ger Nottingham ac es i Ysgol Gyfun Toot Hill yn Birmingham, enw gwych i ysgol! Erbyn hyn, fodd bynnag, rydw i wedi byw'n hirach yng Nghaerhirfryn ar ôl cyrraedd yno fel myfyriwr israddedig ym 1987. Arhosais yno wedi hynny fel cynorthwy-ydd ymchwil a myfyriwr PhD, ac yna fel cymrawd addysgu, darlithydd, uwch-ddarlithydd a bellach yn ddarllenydd mewn prosesu iaith naturiol. Cyn i mi fod yn 10 oed, drwy ddylanwad fy nhad-cu, dechreuais gefnogi Nottingham Forest ac, wrth reswm, nhw yw'r tîm gorau yn y byd o hyd! Bryd hynny, roeddem wedi ennill dau Gwpan Ewrop yn olynol, ac fe wnaethom eu dilyn i Wembley wedi hynny ar gyfer rownd derfynol Cwpan y Gynghrair. O ystyried pa dîm rwy'n ei gefnogi, mae gweithio gydag aelodau eraill tîm prosiect CorCenCC yn llawer o hwyl gan fod un ohonynt yn gefnogwr Derby County (bwww!) ac un arall wedi chwarae i dîm merched Forest (hwrê!). Allwch chi ddyfalu pwy? Os nad ydych yn siŵr, darllenwch yr adrannau 'Cwrdd â'r Tîm' mewn cylchlythyron blaenorol!!

Dechreuais ymddiddori mewn cyfrifiaduron yn yr ysgol yn nyddiau godidog ZX Spectrums a'r BBC Micros. Roedd yn gymysgedd o chwarae gemau a theipio rhaglenni i mewn o gylchgronau, neu eu llwytho drwy ddefnyddio tapiau casét (blant, gofynnwch i'ch rhieni beth yw casét!). I bob pwrpas, ni oedd yn rhedeg labordy cyfrifiaduron yr ysgol ar ran yr athrawon, a chefais BBC Micro fel anrheg pen-blwydd a Nadolig ar y cyd un flwyddyn. Bryd hynny, 64k oedd cof BBC Micros. Pwy yn y byd fyddai angen mwy na hynny?! Erbyn hyn, mae gennyf ebyst sy'n fwy na 64K. Fe benderfynais fynd i Gaerhirfryn am ei bod yn un o ddim ond pum prifysgol yn y DU ar y pryd oedd yn cynnal cwrs gradd cydanrhydedd mewn Cyfrifiadureg a Mathemateg. Yn ogystal, cefais gynnig ysgoloriaeth o £70 ganddynt. Roedd £70 yn llawer o arian ym 1987! Cymerais y cam cyntaf tuag at brosesu iaith naturiol pan ddewisais brosiect israddedig



trydedd blwyddyn yn adeiladu rhaglen parser CYK o dan oruchwyliaeth Roger Garside. Wedi hynny, cefais gynnig swydd ganddo (dim angen cyfweliad bryd hynny!) cyn fy arholiadau terfynol ym 1990. Os neidiwn ymlaen i ddydd Mawrth 24ain Medi 2013, dyna pryd y cefais fy ngwahodd gan Dawn i fynd ar drên (wel, sawl trên i ddweud y gwir) i Gaerdydd ar gyfer cyfarfod i gynnig lansio CorCenCC yn adeilad y Senedd. Mae'r gweddill, ys dywedon nhw, yn hanes. Bûm yn gweithio ar NLP ac yn benodol, tagwyr semantig, yn wreiddiol ar gyfer y Saesneg ym 1990, ond ar gyfer nifer o ieithoedd eraill ers hynny. Felly, mae creu tagiwr semantig ar gyfer y Gymraeg yn y prosiect CorCenCC ar y cyd â Scott Piao wedi bod yn waith diddorol dros ben. Cyn hynny, yr unig eirfa Gymraeg yr oeddwn wedi'i weld oedd yr arwyddion ffyrdd wrth fynd ar fy ngwyliau i Gymru. Wrth gwrs, yr atyniad mawr arall i weithio ar y prosiect yw'r cyfle i gael cacen gri yn rheolaidd! Yng Nghaerdydd mae'r Doctor Who Experience hefyd, felly sut allwn i wrthod?!

+ Cysylltu â ni

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 neu ewch i'n gwefan, sef: www.corcencc.cymru

Mae CorCenCC yn brosiect ymchwil a ariennir gan ESRC/AHRC (Grant Rhif ES/M011348/1). Mae tîm CorCenCC yn cynnwys y Prif Ymchwilydd - Dawn Knight; y Cyd-Ymchwilwyr - Tess Fitzpatrick, Steve Morris, Irena Spasić, Paul Rayson, Enlli Thomas, Mark Stonelake a Jeremy Evas; y Cynorthwywyr Ymchwil - Steven Neale, Jennifer Needs, Mair Rees, Scott Piao a Lowri Williams; y myfyriwr PhD - Vigneshwaran Muralidaran; Ymgynghorwyr - Kevin Donnelly, Kevin Scannell, Laurence Anthony, Tom Cobb, Michael McCarthy a Margaret Deuchar; Grŵp Ymgynghorol y Prosiect - Colin Williams, Karen Corrigan, Llion Jones, Maggie Tallerman, Mair Parry-Jones, Gwen Awbery, Emyr Davies (CBAC-WJEC), Gareth Morlais (Llywodraeth Cymru), Owain Roberts (Llyfrgell Genedlaethol Cymru), Aran Jones (Saysomethingin.com) ac Andrew Hawke (Geiriadur Prifysgol Cymru). Os oes gennych unrhyw sylwadau neu gwestiynau am gynnwys y cylchlythyr hwn, cysylltwch â Dr Dawn Knight: KnightD5@caerdydd.ac.uk