



Cyfarchion gan y Prif Ymchwilydd

Croeso i'r 20fed rhifyn o gylchlythyr CorCenCC. Yn y rhifyn hwn cewch newyddion am ddiwyddiadau diweddar sy'n ymwneud â CorCenCC



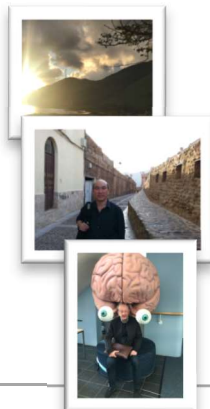
ynghyd ag ambell ddatblygiad cyffrous o ran y prosiect, gan gynnwys y cipolwg cyntaf ar yr offer holi a ddefnyddir i borri, chwilio a

dadansoddi'r corpws. Hefyd, cewch eich cyflwyno i aelod arall o deulu CorCenCC, Scott Piao, a rhoddwn groeso twymgalon i ddau aelod newydd o'r tîm a ffarwelio â Lowri Williams sy wedi gadael y tîm. Gan mai hwn fydd rhifyn olaf eleni, hoffwn fanteisio ar y cyfle i ddymuno Nadolig Llawen IAWN i'n holl ddarllenwyr a chefnogwyr a phob hwyl yn 2019. Edrychaf ymlaen at weld pawb eto ym mis Ionawr. Iechyd da!

Mwynhewch! Dr Dawn Knight

Cynnwys

- T1: Digwyddiadau
- T2: Yr offer holi
- T3: Cipolwg
- T4: Cwrdd â'r tîm
- T5: Hwyl fawr a helo
- T6: Cysylltu â ni

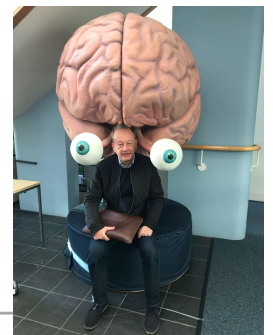


+ Digwyddiadau Ymweld â Phrifysgol Bangor

Yn ddiweddar, bu Dawn Knight, Paul Rayson a Steve Morris ar ymweliad ag aelodau'r tîm ym Mangor a chafwyd cyfle i gwrdd â Delyth Prys, pennaeth yr Uned Technolegau Iaith a Gruff Prys, Uwch Derminolegydd yng Nghanolfan Bedwyr. Roedd hwn yn gyfle gwych i ddysgu mwy am y gwaith arloesol sy'n cael ei wneud yng Nghanolfan Bedwyr mewn nifer o feysydd amrywiol o safoni terminoleg, y geiriadur ar-lein adnabyddus *Ap Geiriaduron* i ddatblygu technolegau testun a lleferydd (gan gynnwys prosiect anhygoel *Lleisiwr* sy'n golygu bod clefion a all fod mewn perygl o golli eu lleisiau yn gallu cadw eu llais i'w ddefnyddio maes o law fel un synthetig digidol personol). Gall darllenwyr ddysgu mwy am y meysydd niferus y mae Canolfan Bedwyr yn weithgar ynddynt trwy'r tri phorth ar-lein y mae modd eu cyrchu trwy'r dolenni canlynol:

- Porth Termau Cenedlaethol Cymru
<http://termau.cymru/>
- Porth Corpora Cenedlaethol Cymru
<http://corpws.cymru/>
- Porth Technolegau Iaith Cenedlaethol Cymru
<http://techiaith.cymru/>

Ar ben hynny, roedd hwn yn gyfle i'r Uned Technolegau Iaith ddysgu mwy am CorCenCC ac archwilio i ffyrdd lle y gallai fod synergedd rhwng ein gwaith yn y dyfodol. Yn ystod yr ymweliad, cafwyd cyfle i ddal i fyny gyda chydweithwyr eraill o Fangor gan gynnwys Kevin Donnelly, Llïon Jones (Cyfarwyddwr Canolfan Bedwyr) ac arweinydd WP4, Enlli Thomas. Cynhaliwyd cyfarfod gyda Manon Jones o'r Ysgol Seicoleg i drafod meysydd posibl lle mae ein hymchwil yn gorgyffwrdd ac i ledu'r gair am weledigaeth a nodau CorCenCC.



Arddangosfa CUROP: Cyflwyno ein Gwaith (Tachwedd 2018)

Wedi imi gwblhau fy lleoliad gwaith haf yn gweithio ar y tasgau WP3, yr oedd yna'n amser i baratoi ar gyfer yr arddangosfa CUROP, digwyddiad lle fyddai cyfle imi adfyfrio ar yr hyn yr oeddwn wedi'i gyflawni a chyflwyno fy ngwaith. Roedd yn gyfle imi arddangos fy nghyfraniad tuag at brosiect CorCenCC drwy greu poster academaidd fy hun (fel y gwelir yn y lluniau) ac i hyrwyddo'r prosiect i'm gyd-fyfyrwyr CUROP. Yn ystod yr arddangosfa, roedd y myfyrwyr a'i chynghorwyr i gyd yn sgwrsio, yn hysbysu ei gilydd a'r cyhoedd o'u prosiectau ymchwil amrywiol, gan roi cyfle i gyflwyno'r prosiect i gynulleidfa ehangach. Roedd yna gryn dipyn o ddiddordeb yn y prosiect a chafodd sawl gwestiwn ei ofyn, rwy'n gobeithio fy mod i wedi eu hateb yn dda! Roedd yn ffordd dda o ddod a'm cyfnod gwaith CUROP i ben, ac yn brofiad gwerthfawr o gyflwyno fy ngwaith a'r prosiect i'r cyhoedd.

Alys Greene




+ Yr Offer Holi

Yn ystod yr wythnosau diwethaf mae offer cychwynnol holi corpws CorCenCC wedi bod yn syrthio i'w lle, gan ddarparu porth i ddefnyddwyr gael mynediad i ddata terfynol y corpws. Mae'r offer yn cael eu datblygu fel rhan o WP5, sy'n canolbwyntio ar y seilwaith sy'n ofynnol i adeiladu a chynnal y data, a sicrhau bod pobl yn gallu gwneud yn fawr ohono pan fydd yn barod. Mae hon yn agwedd arbennig o gyffrous ar ddatblygiad technegol y prosiect, wrth i ni ddechrau llunio allbwn gweladwy CorCenCC a gwireddu dull o ddarlunio, holi a dadansoddi ein set ddata gyfoes o Gymru.

Mae ein gwaith cynnar wedi canolbwyntio ar rai o'r prif nodweddion swyddogaethol sy'n gysylltiedig ag offer dadansoddi a holi corpws, gan gynnwys gair allweddol yn ei gyd-destun (KWIC), llinellau mynegeiriau, rhestrau amledd, dadansoddiad n-gram, a dadansoddiad cydleoli. Wrth gwrs, mae ein dull egwyddorol o gasglu data yn golygu bod modd hidlo canlyniadau holi hefyd yn ôl yr amrywiol fetadata rydym wedi bod yn ei gasglu wrth fynd ati i grynhoi'r data, a bydd hynny'n ddiamau yn bwrw goleuni diddorol ar sut

mae'r Gymraeg yn cael ei defnyddio mewn gwahanol gyd-destunau! Yn naturiol, mae datblygiad yr offer yn cael ei lywio gan arolwg diweddar a gynhaliwyd gennym ar offer sydd eisoes yn bodoli i ddadansoddi a holi'r corpws, y mae ymchwilwyr, ymarferwyr a ieithyddion yn defnyddio ystod eang ohonynt at amrywiaeth o ddibenion. Mae'r adborth a gawsom am y pethau sy'n gweithio'n dda wedi bod yn ddiddorol iawn, ac rydyn ni'n edrych ymlaen at weld sut mae pobl yn defnyddio'r gwahanol nodweddion rydyn ni'n eu cynnwys!



Simple Query > Results

Word: -- | Lemma: 'bod'

POS: -- | Mutation: sm

Filtered by: --

Browse Metadata

Corpus query tools

Total results: 86 of 14876 (0.58%)

No.	Filename	Keyword
1	Electronic (1/4)	. Ar yr un pryd newidiodd Cymru mewn cenhedlaeth neu ddwy o
2	Electronic (1/4)	newidiodd Cymru mewn cenhedlaeth neu ddwy o fod yn wlad Gatholig i
3	Electronic (1/4)	ac am hunanlywodraeth ac erbyn diwedd y 19eg ganrif roedd mudiad Cymru
4	Electronic (1/4)	y 19eg ganrif roedd mudiad Cymru Fydd ar ei anterth . Cymysg
5	Electronic (1/4)	r 1980au , mae Cymru heddiw 'n meddu Cynulliad Cenedlaethol ac ymddengys
6	Electronic (1/4)	Fynwy " , anomaledd a barhaodd hyd yr 20fed ganrif er na
7	Electronic (2/4)	Tynged yr Iaith , sef darlith Radio BBC Cymru : y canlyniad
8	Electronic (2/4)	. Roedd yn gredwr cryf yn y traddodiad Ewropeaidd , a gwelodd
9	Electronic (3/4)	yn annerbyniol i'r Cymry Cymraeg ac i'r di-Gymraeg hefyd am
10	Electronic (3/4)	y Deyrnas Unedig , ond bydd hyn yn dod i ben pan
11	Electronic (3/4)	Disgrifiodd Cymdeithas yr Iaith y penderfyniad fel " anghredadwy " gan rybuddio
12	Electronic (3/4)	bodoli . Y gwir yw , erbyn 2015 , mae 'n bosib
13	Electronic (3/4)	. Ymateb Gweinidog Treftadaeth Cymru , Alun Ffred Jones AC , oedd
14	Electronic (3/4)	" . Ychwanegodd nad oedd unrhyw drafodaeth wedi bod , a 'i
15	Electronic (3/4)	dal heb gael gwybod yn swyddogol am y penderfyniad . Dywedodd ei
16	Electronic (3/4)	yr Iaith Gymraeg . Dywedodd Cadeirydd y Bwrdd , Meri Huws ,
17	Electronic (3/4)	2010 dywedodd cyn-Uwch Gyfarwyddwr S4C , Geraint Stanley Jones (a
18	Electronic (3/4)	(a fu yn y swydd o 1989 tan 1994) na
19	Electronic (3/4)	'r sianel yn dechrau brwydro yn erbyn y llywodraeth . Dywedodd hefyd
20	Electronic (3/4)	Awdurdod S4C hefyd gan yr aelod seneddol Caidwadol Alun Caimis a ddywedodd
21	Electronic (3/4)	Cymru . Mae'n Ddiabed . Dywedodd Ms Diabed a bod yn fela

Mynegeirydd CorCenCC

Yn ystod y misoedd nesaf, byddwn ni'n ymhelaethu ar y gwaith hyd yma i gynnwys cynifer o nodweddion defnyddiol â phosibl, fel bod defnyddwyr yn gallu cael yr holl wybodaeth sydd ei hangen arnyn nhw am Gymraeg cyfoes trwy CorCenCC. Byddwn ni hefyd yn dechrau cynllunio ar gyfer integreiddio ein pecyn offer addysgeg - sy'n cael ei ddatblygu fel rhan o WP4 - er mwyn galluogi athrawon a dysgwyr i wneud y defnydd gorau posibl o'r data ar gyfer eu cynlluniau gwersi a'u sesiynau astudio eu hunain. Efallai mai'r elfen fwyaf cyffrous yw y byddwn hefyd yn dechrau'r dasg o fwydo'r data terfynol a gasglwyd gan y tîm WP1 i mewn i'r offer, a dyna lle byddwn ni'n gweld y set ddata derfynol yn dechrau ymffurfio!

Ciplowg: beth sy'n digwydd i'r data unwaith y caiff ei gasglu?

Os ydych wedi bod yn dilyn ein cylchlythyr, fe welwch ein diweddariadau ar ein cynnydd o gasglu data. Nod CorCenCC yw ffurfio corpws o 10 miliwn o eiriau o ddata llafar, ysgrifenedig ac e-iaith. Rydym wedi mabwysiadu sawl techneg i gasglu data o'r fath, o ddefnyddio dulliau awtomatig megis "crafwr gwe", i fynyachu digwyddiadau, cyfarfod dilynwyr y prosiect a recordio eich sgysia. Ond unwaith y bydd y data hwn wedi'i gasglu, beth sy'n digwydd iddo?

Unwaith y bydd aelod o'n tîm ymchwil wedi casglu data, mae'n dilyn sawl cam prosesu cyn y gellir ei rhoi i mewn i'r corpws terfynol y bydd gennych chi, y cyhoedd, mynediad iddo. Ar gyfer data llafar, y cam cyntaf yw sicrhau ei fod wedi'i logio. Rydym yn nodi lle cafodd ei recordio, pwy yw'r cyfranogwyr o'r recordiad, a sicrhau bod ansawdd y recordiad yn glir. Yna caiff y recordiad ei storio'n ddiogel ar ein gweinyddwyr diogel. Gan fod CorCenCC yn gorpws testun, ac nid un llafar, mae angen trosi'r recordiad mewn i ffurf ysgrifenedig. Yn yr achos hwn, mae gennym dîm o drawsgrifwyr CorCenCC sy'n cynhyrchu ffurf ysgrifenedig o'ch sgysia. Diolch yn fawr i'n trawsgrifwyr sy'n parhau i weithio'n galed ar y dasg hon! Cyn y gellir cynnwys data trawsgrifiadau, ysgrifenedig ac e-iaith fel rhan o gorpws terfynol CorCenCC, mae gan ein tîm ymchwil y dasg bwysig o wirio ansawdd y data. Mae hyn yn golygu sicrhau bod yr holl wybodaeth personol wedi'i dileu. Unwaith y caiff ansawdd y data ei wirio, caiff ei lwytho i fyny i'n gweinyddwr ac mae'n barod i'w fewnosod yn y corpws terfynol. Yn dibynnu ar faint y data, mae'r dasg o wirio yn aml yn cymryd peth amser i'w gwblhau. Serch hynny, mae'n eithaf cyffrous gweld y corpws yn llenwi gyda'ch Cymraeg chi.

Rhowch eich Cymraeg i ni!

Ydych chi'n defnyddio WhatsApp i decstio yn Gymraeg? Dyn ni angen eich help! Hoffen ni gynnwys enghreifftiau o negeseuon testun yn y corpws - maen nhw'n ffordd eitha unigryw o gyfathrebu! Allech chi anfon enghreifftiau o'ch negeseuon aton ni? Mae'n hawdd iawn i'w wneud. **Anfonwch neges WhatsApp at +44 7542 348512** gan ddweud ai iPhone neu ffôn Android yr ydych yn ei ddefnyddio ac fe wnawn ni esbonio beth i'w wneud nesaf.

+ Cwrdd â'r tîm: Scott Piao, cyn GY ym Mhrifysgol Caerhirfryn ac ymgynghorydd prosiect erbyn hyn

Yn gyntaf oll, mae wedi bod yn brofiad gwyach i mi fod yn rhan o ddatblygiad prosiect CorCenCC, a gweithio gyda thîm rhagorol y prosiect i wireddu'r syniad mawr hwn. Cafodd fy niddordeb a'm profiad ym maes dadansoddi data iaith a datblygu offer meddalwedd at ddiben o'r fath fodd i fyw yn y prosiect hwn, ac mae wedi bod yn bleser o'r mwyaf i mi gael cydweithio ag aelodau'r tîm i ddatblygu offer ar gyfer yr iaith Gymraeg, un o'r prif ieithoedd imi weithio gyda nhw erioed.

Mae fy angerdd ynghylch datblygu offer a



systemau ar gyfer dadansoddi gwybodaeth ar sail iaith naturiol yn cysylltu nôl â'm cyfnod fel myfyriwr israddedig yn y brifysgol yn Tsieina, pan ges i gyfle i astudio dau brif bwnc, cyfrifiadureg a ieithoedd. Wrth ddod i gysylltiad â chyfrifiaduron am y tro cyntaf, ces i fy hudod mewn dim o dro gan raglennu cyfrifiadurol. Bryd hynny, iaith raglennu BASIC (sy'n hynafol heddiw) oedd y brif iaith, ac rwy'n dal i gofio'r cyffro pan welais i fy rhaglen BASIC yn argraffu calendr syml am y tro cyntaf ar ddarn o bapur print gydag ymylon darniog yn llawn tyllau (rhywbeth arall sy'n hynafol erbyn heddiw). Wrth i amser fynd heibio, mae BASIC yn datblygu'n C, ac yna eto'n Java, Python, a phob math o ieithoedd cyfrifiadurol newydd sbon, ond mae fy mrwdfrydedd ynghylch datblygu systemau meddalwedd wedi parhau hyd heddiw. Ochr yn ochr â'r cyfrifiaduwr, mae iaith wedi bod yn elfen arall o'm diddordebau craidd. Gan fy mod i wedi cael hyfforddiant ym maes

ieithyddiaeth, rwyf wedi mwynhau tyrchu mewn gwybodaeth systematig am ieithoedd a ieithyddiaeth, ac am rai blynyddoedd bues i'n darlithio ar gwrs ieithyddiaeth. Wrth ddod i Brifysgol Lancaster yn 1996, roeddwn i mor gyffrous i gael hyd i fyw newydd yn UCREL, dan arweiniad Geoff Leech ar y pryd, lle gallwn i gyfuno fy sgiliau a'm gwybodaeth am ieithoedd a chyfrifiadureg, ac fe gyflawnais i brosiect PhD mewn Ieithyddiaeth Corpws ym Mhrifysgol Lancaster.

Wrth gwrs, daliodd fy niddordebau ymchwil i esblygu ac ehangu. Trodd fy llwybr ymchwil i gyfeiriad arall eto yn sgîl fy swydd gyntaf yn y Deyrnas Unedig, ym Mhrifysgol Sheffield. Yn y Grŵp Prosesu Iaith Naturiol (NLP) yno, fe ges i gyfle i ddysgu llawer am faes ymchwil NLP yn ei grynswth, a datblygais diddordeb pendant yn y maes hwn. Cymerodd fy ngwybodaeth a'm profiad ym maes NLP gam arall ymlaen yn ystod fy ngwaith yn y Ganolfan Genedlaethol ar gyfer Cloddio mewn Testun (NaCTeM) ym Mhrifysgol Manceinion. Yn ystod cyfres o brosiectau y bues i'n gweithio arny'n nhw yn ystod y 18 mlynedd diwethaf, rwyf bellach wedi datblygu diddordebau ymchwil eang, sy'n cwmpasu NLP, Cloddio mewn Testun,



Ieithyddiaeth Corpws, Cyfrifiadureg Gymdeithasol a Gwyddor Data, ond mae'r cyfan wedi'i wreiddio mewn dadansoddi data iaith.

I droi'n ôl at brosiect CorCenCC, mae datblygu'r tagiwr semantig ar gyfer yr Iaith Gymraeg yn parhau ag ymdrechion blynyddoedd lawer i ddatblygu offer dadansoddi semantig ar gyfer cynifer o ieithoedd â phosibl, ar sail yr offeryn

Saesneg a gychwynnwyd gan Paul Rayson et al. Hyd yma, mae'r tagiwr Cymraeg wedi casglu a chrynhoi'r adnoddau geirfaol a iaith mwyaf ymhlith y tagwyr semantig heblaw Saesneg, a hynny yn sgîl ymdrechion cydweithredol tîm y prosiect. Wrth gwrs, oherwydd rhai o nodweddion unigryw'r iaith Gymraeg a'r diffyg profiad o brosesu'r iaith Gymraeg yn gyffredinol, mae'n dal yn her sylweddol sicrhau bod y tagwyr Cymraeg yn rhedeg yn gywir. Mae'n arbennig o anodd sicrhau cyfatebiaeth ddibynadwy rhwng strwythur cystrawennol unedau geirfaol sy'n cael eu cyfieithu i'r Gymraeg a'r Saesneg, ac felly mae'n her fawr gwneud defnydd llawn o'r adnoddau semantig Saesneg sydd eisoes yn bodoli i greu elfennau cyfatebol yn Gymraeg. Un posibilrwydd yw cymhwyso technegau dysgu peiriant dwfn i ddadelfennu categorïau semantig geiriau ac ymadroddion Cymraeg ar sail data corpws mawr.

Ar 1af Awst, cychwynnais ar fenter newydd, sef bod yn ddarlithydd academiaidd (ydw, rwy'n dechrau'n hwyr, ond rwy'n gobeithio y bydda i'n llwyddo'n hwyr hefyd!), ac felly mae'n rhaid i mi gamu i lawr o'm rôl fel uwch gydymaith ymchwil ym Mhrosiect CorCenCC. Ond fydda i ddim yn mynd yn bell o gwbl, a bydda i'n cadw mewn cysylltiad agos â'r prosiect hwn ac yn cyfrannu ato lle bynnag y galla i er mwyn ei helpu i lwyddo.

+ Hwyl fawr a helo

Gwaetha'r modd, rydym wedi gorfod ffarwelio â Chynorthwydd Gweinyddol Prifysgol, Lowri Williams, sydd wedi derbyn swydd barhaol fel swyddog ystadegol yn y Swyddfa Ystadegau Gwladol. Mae Lowri wedi bod yn gaffaeliad mawr i'r tîm a gwelwn ei heisiau'n fawr. Dymunwn bob lwc iddi yn ei swydd newydd a gobeithiwn y bydd yn dal i ymwneud â phrosiect CorCenCC mewn rhyw ffordd neu'i gilydd. Mae'n dda gennym gyhoeddi bod Alys Greene, myfyrwraig Israddedig ym Mhrifysgol Caerdydd ar hyn o bryd a weithiodd fel ymchwilydd ar leoliad CUROP dros yr haf, wedi ymuno fel cynorthwydd gweinyddol. Croeso yn ôl i'r tîm Alys – mae'n wych dy fod yn gweithio i dîm CorCenCC unwaith eto!

Hefyd, mae'n dda gennym gyhoeddi ein bod wedi recriwtio cynorthwydd ymchwil newydd i weithio ar WP3 ym Mhrifysgol Caerhirfryn, Ignatius Ezeani. Mae Ignatius yn Gydymaith Ymchwil yng Nghanolfan Ymchwil UCREL, Prifysgol Caerhirfryn. Ar hyn o bryd, mae ei ddiddordebau ymchwil yn ymwneud â datblygu fframweithiau cadarn i addasu modelau a thechnegau presennol NLP ar gyfer ymchwil iaith adnoddau prin. Mae ganddo ddiddordeb arbennig yn y math o haniaethau ystyr a pherthnasau semantig sy'n cael eu dal gan fodelau mewnblannu dwfn a hyfforddir yn aml gan feintiau enfawr o ddata oddi wrth ieithoedd â llawer o adnoddau a sut i gymhwyso'r rhain ar gyfer ieithoedd adnoddau prin. Ar ben hynny, mae ganddo ddiddordeb cyffredinol mewn dylunio a datblygu dysgu trwy beiriannau a modelau niwral dwfn yn ogystal â chymhwyso'r rhain, nid yn unig i NLP, ond hefyd i faes ehangach gwyddor data. Ar hyn o bryd, mae Ignatius wrthi'n edrych ar ddulliau effeithlon o wella manwl gywirdeb a dibynadwyedd y Tagiwr Semantig Cymraeg. Croeso mawr i ti Ignatius!



Alys Greene



Lowri Williams



Ignatius Ezeani

+ 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



Arts & Humanities
Research Council

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-Ymchwilywyr - Tess Fitzpatrick, Steve Morris, Irena Spasić, Paul Rayson, Enlli Thomas, Alex Lovell a Jonathan Morris; y Cynorthwywyr Ymchwil - Steven Neale, Jennifer Needs, Mair Rees, Scott Piao a Lowri Williams; y myfyrwyr PhD - Vigneshwaran Muralidaran a Bethan Tovey; 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