Cylchlythyr CorCenCC

Rhifyn 8: Tachwedd 2016

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Cyfarchion gan y Prif Ymchwilydd

Felly dyma ni – wythfed cylchlythyr CorCenCC, sy'n nodi diwedd naw mis o waith ar y prosiect. Yn sicr, mae amser yn hedfan pan 'dych chi'n cael hwyl! Fis diwethaf roedden ni wedi'ch hysbysu bod yr ap torfoli bron â bod yn barod. Diolch i nifer o wirfoddolwyr, erbyn hyn mae hwnnw wedi cael ei dreialu'n drwyadl ac mae'r newidiadau/cyffyrddiadau bach olaf wrthi'n cael eu gwneud nawr. Gobeithio bydd yr ap ar gael yn siop ap Apple yn fuan iawn, felly cadwch eich llygaid ar agor (peidiwch â phoeni, bydd fersiwn Android ar ei ffordd yn 2017 hefyd)! Mae cwblhau'r ap yn cyd-fynd yn agos ag agwedd fwyaf llafurus y prosiect: casglu data. Fel y gwyddoch, rydyn ni'n bwriadu casglu o leiaf 10 miliwn o eiriau o Gymraeg (fel y mae hi'n cael ei defnyddio) erbyn diwedd y prosiect. Er mwyn rhoi gwell syniad i chi o'r gwahanol fathau o ddata sy'n mynd i gael eu cynnwys, a natur y cyfranogwyr yr hoffen ni eu denu i gyfrannu, bydd y tri chylchlythyr nesaf yn cynnig trosolwg o'r 'fframwaith samplo' h.y. cynlluniau casglu data ar gyfer e-iaith, iaith lafar ac iaith ysgrifenedig yn eu tro. Gobeithio bydd hyn yn helpu i roi darlun ehangach o'r prosiect i chi ac yn eich temtio i gymryd rhan a chyfrannu'ch Cymraeg chi! Hefyd, yn y rhifyn hwn, cewch newyddion diweddaraf y tîm ac unwaith eto byddwn yn cyflwyno aelod arall o'r tîm (Laurence Anthony).

Hoffen ni eich hysbysu mai hwn fydd cylchlythyr olaf y flwyddyn. O hyn allan, byddwn yn cynhyrchu'r cylchlythyr bob yn ail fis (felly gallwch ddisgwyl yr un nesaf rywbryd ym mis Ionawr). Bydd y cylchlythyrau yn parhau i roi'r newyddion cyfredol, sôn am ein cynnydd a rhoi gwybodaeth arall am y prosiect – ond mewn fformat mwy swmpus, llawn hwyl. Mae'n teimlo'n amserol felly i ddymuno Cyfarchion y Tymor i chi i gyd – rydyn ni (tîm CorCenCC) yn gobeithio y cewch chi Nadolig bendigedig ac edrychwn ymlaen at gysylltu â chi eto yn y Flwyddyn Newydd.

Mwynhewch! Dr Dawn Knight (Prifysgol Caerdydd)

Newyddion

Ym mis Chwefror 2017 bydd Tess Fitzpatrick, cyd-ymchwilydd ar brosiect CorCenCC, yn symud o un o'n prifysgolion partner i un arall! Bydd yn gadael Caerdydd i dderbyn swydd Pennaeth Saesneg (Iaith) ac Ieithyddiaeth Gymhwysol ym Mhrifysgol Abertawe lle bydd yn ymuno ag aelodau tîm CorCenCC Jenny Needs, Mair Rees, Mark Stonelake a Steve Morris. Meddai Tess *"Er bydda i'n gweld eisiau gweithio'n ddyddiol gyda'r cydweithwyr ardderchog yn Iaith a Chyfathrebu yng Nghaerdydd, bydd y ffaith bod tîm y prosiect yn pontio'r ddwy brifysgol yn gwneud y symud gymaint yn haws. Rwy'n edrych ymlaen at yr heriau a'r cyfleoedd a ddaw yn sgil y rôl newydd ac at weithio gyda chydweithwyr a chyd-destunau newydd. Bydd y prosiect yn allweddol o ran sefydlu ac adeiladu ar y tir cyffredin rhwng gweithgareddau Cymraeg ac Ieithyddiaeth Gymhwysol yn Abertawe wrth i ni barhau i greu ac atgyfnerthu cysylltiadau rhwng partneriaid academaidd a chymunedol y prosiect"*.

Rho dy Gymraeg i ni: Ffocws ar e-iaith

Fel y gwyddoch, un o amcanion allweddol prosiect CorCenCC yw creu corpws sy'n gytbwys ac sy'n adlewyrchu'r holl ffyrdd y mae'r Gymraeg yn cael ei defnyddio o ddydd i ddydd yn y byd 'go iawn'. Mae hyn yn golygu y bydd y corpws yn cynnwys iaith o ystod o fathau gwahanol, yn trafod pynciau gwahanol, gan wahanol gyfranwyr o bob cefndir. Bydd dwy filiwn o'r deg miliwn o eiriau y gobeithiwn eu casglu yn dod o ffynonellau 'e-iaith'. Yn y bôn,

mae'r rhain yn ddulliau cyfathrebu sy'n cael eu 'geni' yn ddigidol – e.e. yr iaith sy'n cael ei theipio i mewn i'ch ffôn, cyfrifiadur neu unrhyw ddyfais electronig arall. Wrth i'r byd ddod yn fwyfwy digidol mae'n bwysig bod y corpws yn dangos ac yn adlewyrchu sut mae'r iaith yn cael ei defnyddio yn y ffordd hon.

Canolbwyntion ni ar gasglu data e-iaith yn ôl yn 2013, wrth i ni beilota'r syniad o CorCenCC yn y lle cyntaf, ac felly mae gennym brofiad o recriwtio e-gyfranwyr a chasglu eu data. Bryd hynny, llwyddon ni i adeiladu corpws bach gyda thros 500 mil o eiriau gan 72 o bobl wahanol. Ar gyfer y CorCenCC 'go iawn' rydyn ni'n bwriadu ymestyn hyn trwy gasglu tua 600 mil o eiriau yr un o flogiau a gwefannau, a 400 mil o eiriau yr un o ebyst a negeseuon testun gan gannoedd o ddefnyddwyr y Gymraeg. Yn wreiddiol, roedden ni hefyd yn gobeithio samplo trydariadau, gan ein bod yn deall y caiff Trydar ei ddefnyddio'n helaeth yng nghyd-destun yr iaith Gymraeg. Fodd bynnag, mae rheolau a thelerau Trydar yn cyfyngu ar faint o drydariadau a gaiff eu rhannu â thrydydd partïon. Gan y bydd holl allbynnau CorCenCC ar gael i aelodau'r cyhoedd, 'dyw cyfyngiadau o'r fath ddim yn cyd-fynd ag ethos CorCenCC, ac felly yn anffodus fydd hi ddim yn bosibl cynnwys trydariadau yn ein corpws.

Fel canllaw cychwynnol ar gyfer samplo'r data, rydyn ni wedi penderfynu casglu'r e-iaith gyhoeddus (h.y. gwefannau a blogiau) yn ôl 6 dosbarthiad thematig bras. Mae'r dosbarthiadau hyn wedi'u seilio ar system ddosbarthu debyg a ddefnyddiwyd ar gyfer corpws e-iaith arall: CANELC (the Cambridge and Nottingham E-Language Corpus – Dawn oedd y cynorthwyydd ymchwil a gasglodd y data ar gyfer y corpws hwn yn ôl yn 2009/2010), ond mae rhai o'r categorïau wedi'u haddasu er mwyn gwneud i'r system adlewyrchu cyd-destun y Gymraeg yn well. Mae'r dosbarthiadau'n amrywio o themâu ffurfiol yn y byd cyhoeddus (e.e. newyddion, y cyfryngau a materion cyfoes) i themâu sy'n fwy anffurfiol a phersonol (e.e. bod yn rhiant a bywyd teuluol).

Yn achos y mathau o e-iaith sy'n fwy preifat (h.y. ebyst a negeseuon testun), fyddwn ni ddim yn gwybod dim byd am y data cyn i ni eu derbyn, wrth gwrs, felly yn hytrach na samplo'r mathau hyn o ddata ar sail themâu, byddwn ni'n targedu cyfranwyr ar sail pwrpas y cyfathrebu, hynny yw a ydyn nhw'n cyfathrebu am faterion personol neu faterion busnes. Mae hyn yn cyd-fynd, fwy neu lai, â'r fframwaith samplo arfaethedig ar gyfer data llafar (cewch ragor ynglŷn â hyn yn y rhifyn nesaf!) yn yr ystyr bod negeseuon busnes yn fwy tebygol o fod yn Drafodol neu'n Broffesiynol eu naws, tra bod negeseuon personol yn fwy tebygol o gael eu categoreiddio fel iaith Gymdeithasu neu Breifat. Mae'n debyg y cawn lai o negeseuon testun electronig byr gan fusnesau, tra bod digonedd o ebyst busnes, ac felly mae targedau geiriau ein fframwaith samplo wedi'u ffurfio gyda hyn mewn golwg. Mae'r fframwaith samplo ar gyfer data e-iaith isod. Eto, dim ond canllaw yw hwn ar gyfer yr hyn yr hoffen ni ei gasglu – 'delfryd'. Mewn gwirionedd mae'n debyg y bydd dosbarthiad y data yn CorCenCC yn eitha gwahanol o'i gymharu â hwn, ond mae'n gweithredu fel man cychwyn/sail ddefnyddiol i ni adeiladu arno. Cewch gip arno ac, os hoffech chi gyfrannu unrhyw rai o'r mathau hyn o ddata iaith i'r corpws, cysylltwch â ni!

Thema/Pwnc / *Theme/Topic*	**%**	**Geiriau /*Words***
Blog	**30%**	**600,000**
A: Newyddion, Y Cyfryngau a Materion Cyfoes / *News, Media and Current Affairs*, Gwleidyddiaeth / *Politics*, Busnes a Chyllid / *Business and Finance*, Y Tywydd a'r Amgylchedd / *Weather and the Environment*, Siopa Ar-lein / *Online Shopping*	5%	100,000
B: Crefydd / *Religion*, Yr Iaith / *Language*, Diwylliant, Llenyddiaeth a'r Celfyddydau / *Culture, Literature and the Arts*, Addysgu, Academia ac Addysg / *Teaching, Academia and Education*	5%	100,000
C: Technoleg, Cyfrifiaduron a Chwarae Gemau Cyfrifiadurol / *Technology, Computers and Gaming*, Ffasiwn a Harddwch / *Fashion and Beauty*, Hobïau a Difyrrwch / *Hobbies and Pastimes*, Teithio / *Travel*, Coginio / *Cookery*	5%	100,000
D: Cerddoriaeth / *Music*, Chwaraeon / *Sport*, Perfformiadau byw a Digwyddiadau / *Gigs and Events*	5%	100,000
E: Hynt a Helynt Pobl Enwog / *Celebrity news and gossip*, Teledu a Ffilm / *TV and Film*, Hiwmor / *Humour*	5%	100,000
F: Bod yn Rhiant a Bywyd Teuluol / *Parenting and Family Life*, Iechyd a Lles / *Health and Wellbeing*, Bywyd Personol a Phob Dydd / *Personal and Daily Life*	5%	100,000

Gwefan / *Website*	30%	600,000
A: Newyddion, Y Cyfryngau a Materion Cyfoes / *News, Media and Current Affairs*, Gwleidyddiaeth / *Politics*, Busnes a Chyllid / *Business and Finance*, Y Tywydd a'r Amgylchedd / *Weather and the Environment*, Siopa Ar-lein / *Online Shopping*	5%	100,000
B: Crefydd / *Religion*, Yr Iaith / *Language*, Diwylliant, Llenyddiaeth a'r Celfyddydau / *Culture, Literature and the Arts*, Addysgu, Academia ac Addysg / *Teaching, Academia and Education*	5%	100,000
C: Technoleg, Cyfrifiaduron a Chwarae Gemau Cyfrifiadurol / *Technology, Computers and Gaming*, Ffasiwn a Harddwch / *Fashion and Beauty*, Hobïau a Difyrrwch / *Hobbies and Pastimes*, Teithio / *Travel*, Coginio / *Cookery*	5%	100,000
D: Cerddoriaeth / *Music*, Chwaraeon / *Sport*, Perfformiadau byw a Digwyddiadau / *Gigs and Events*	5%	100,000
E: Hynt a Helynt Pobl Enwog / *Celebrity news and gossip*, Teledu a Ffilm / *TV and Film*, Hiwmor / *Humour*	5%	100,000
F: Bod yn Rhiant a Bywyd Teuluol / *Parenting and Family Life*, Iechyd a Lles / *Health and Wellbeing*, Bywyd Personol a Phob Dydd / *Personal and Daily Life*	5%	100,000

Ebost / *Email*		20%	400,000
Proffesiynol *Professional*	e.e. ebost i gadarnhau amser cyfarfod *e.g. an email to confirm a meeting*	13.4%	268,000
Personol *Personal*	e.e. ebost sy'n rhannu newyddion da *e.g. an email to share good news*	6.6%	132,000

Negeseuon Testun Electronig Byr / *Short Electronic Text Messages*		20%	400,000
Proffesiynol *Professional*	e.e. Neges sydd wedi ei hanfon gan ysgol sy'n darparu gwybodaeth ynghylch noswaith rhieni / *e.g. a message sent by a school providing details of a parents' evening*	6.6%	132,000
Personol *Personal*	e.e. neges ynghylch cwrdd â ffrind am goffi *e.g. a message regarding meeting a friend for coffee*	13.4%	268,000
		100%	**2,000,000**

Cwrdd â'r tîm

Bob mis byddwn yn rhoi sylw i aelod gwahanol o'r tîm CorCenCC estynedig yn ein cylchlythyr. Bydd hyn yn rhoi cyfle i bawb ddweud ychydig wrthoch chi am eu cefndir; beth maen nhw am ei weld o CorCenCC a sut maen nhw'n credu y gallai gyfrannu at eu gwaith eu hunain, neu yn fwy cyffredinol, at waith eraill yng Nghymru. Y mis yma mae'r sbotolau ar Yr Athro Laurence Anthony, sy'n aelod o Grŵp Ymgynghorol Prosiect CorCenCC, ac sy'n gweithio ym Mhrifysgol Waseda.

Proffil: Laurence Anthony

Pan gefais fy ngwahodd i ysgrifennu cyflwyniad ar fy ngwaith a'm cysylltiad â phrosiect CorCenCC, fy ngreddf gyntaf oedd mynd yn ôl i'r holl gyflwyniadau "Cwrdd â'r tîm" blaenorol i gael syniad am eu harddull, eu hyd, a'r pynciau a ystyriwyd ynddynt. Gwelais yn fuan fod aelodau eraill o'r tîm wedi defnyddio, ar gyfartaledd, 268.8 o ffurfiau geiriol gwahanol yn eu cyflwyniadau, a bod y cyflwyniadau ar gyfartaledd yn 538.4 o eiriau o hyd. Gwelais hefyd fod "Welsh", "language", "resources", "developing", "my" ac "I" yn rhai o eiriau allweddol y cyflwyniadau. Felly, yn awr mae modd i mi fwrw ati a dechrau ysgrifennu! Ie... er gwell neu er gwaeth, dyna fi... ieithydd corpws!

Felly, gadewch i mi gyflwyno fy nghefndir ac egluro fy nghysylltiad â phrosiect CorCenCC. Cefais fy magu yn Huddersfield, yn fab i fam o Gaerdydd a thad o Lundain. Pan oeddwn yn

fachgen, roeddwn yn breuddwydio – yn dyheu - am fod yn astroffisegydd. Roeddwn yn dwlu ar y syniad gymaint, darllenais yr holl lyfrau gan ffisegwyr fel Stephen Hawking a Richard Feynman (es i mor bell â rhoi'r enw Richard i'm mab!), dysgais sut i adeiladu cylchedau electronig ac i raglennu cyfrifiaduron, a hyd yn oed mynd i'r brifysgol i astudio'r pwnc. Ond, wrth astudio yn Athrofa Gwyddoniaeth a Thechnoleg Prifysgol Manceinion (UMIST), cymerodd fy llwybr gyrfaol dro dramatig wrth fynd ar daith i Siapan a chyfarfod llawer o wyddonwyr a pheirianwyr yno a oedd yn ei chael yn anodd cyfathrebu yn yr iaith dramor yna - Saesneg. Penderfynais bryd hynny, wedi graddio, y byddwn yn symud i Siapan a chanolbwyntio yn fy ngyrfa ar ddysgu cyfathrebu gwyddonol, a datblygu adnoddau addysgol ac ymchwil i helpu pobl i ddysgu a dadansoddi iaith.

Bum mlynedd ar hugain yn ddiweddarach, rwy'n gwneud yr un math o waith. Ond mae un peth, yn amlwg iawn, wedi newid dros y cyfnod hwnnw. Po hwyaf fy nghyfnod y tu allan i'r DU, fwyaf rydw i wedi gwerthfawrogi'r gwerth, y pwysigrwydd, a'r mewnwelediadau a geir o nabod mwy nag un iaith. Felly, dyma ganolbwyntio llawer o'm hamser ar ddatblygu offer sy'n cefnogi ymchwil, addysgu, a dysgu ieithoedd lu. Yn wir, un o'm rhaglenni meddalwedd fwyaf poblogaidd yw offeryn dadansoddi corpws rhadwedd (*freeware*) o'r enw *AntConc* sy'n cael ei ddefnyddio gan bobl mewn mwy na 140 o wledydd ar draws y byd. O ystyried hyn, efallai mai hawdd yw deall pam y neidiais at y cyfle i fod yn ymgynghorydd ar brosiect CorCenCC, lle y gallaf gefnogi tîm CorCenCC wrth iddynt greu offer dadansoddi ymchwil ac adnoddau dysgu ar gyfer y Gymraeg.

O'm safbwynt fel ieithydd corpws, mae prosiect CorCenCC yn arloesol mewn sawl ffordd. Er enghraifft, er mwyn casglu data Cymraeg dilys, bydd y prosiect yn defnyddio dulliau torfol sydd ar flaen y gad, a fydd yn ein galluogi i weld sut mae'r iaith yn cael ei defnyddio mewn sefyllfaoedd go iawn, pob dydd. Hefyd, gan fod y prosiect ar raddfa mor eang, bydd yn rhaid storio'r data mewn ffordd unigryw fel y gall pobl gael mynediad at y nifer enfawr o ganlyniadau yn gyflym ac yn hawdd. Mae'r prosiect hefyd yn torri tir newydd o ran y ffordd mae'n mynd ati i gysylltu casglu data iaith â datblygu offer a deunyddiau dysgu ac addysgu ymarferol. Felly, mae'n anrhydedd go iawn i fod yn rhan o'r prosiect ac i helpu mwy o bobl i ddeall, dysgu, ac addysgu'r Gymraeg.

Nawr... dyma'r cyflwyniad wedi cyrraedd 577 o eiriau o hyd, felly mae'n debyg y dylwn roi'r gorau iddi!

Laurence Anthony

CorCenCC ar-lein

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: http://sites.cardiff.ac.uk/corcencc/

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 Gareth Watkins; **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