Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

CyTag
Rule-based tagging for Welsh

Cynnwys

T2: Newyddion a digwyddiadau

T4: Diweddariadau cyffredinol

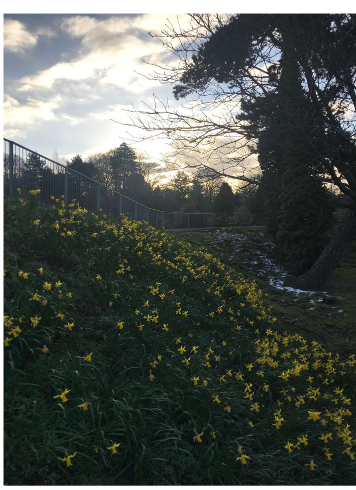
T5: Cwrdd â'r tîm

T6: Cysylltu â ni

Cyfarchion gan y Prif Ymchwilydd

Croeso i'r gwanwyn (gobeithio ein bod wedi gweld cefn ar yr eira...)! Mae mis Mawrth yma o'r diwedd: mae'r cennin Pedr yn blodeuo, mae'r ŵyn yn brefu ac mae'r tîm yn dal i weithio'n galed yn casglu data ac yn gosod y seiliau ar gyfer Corpws Cenedlaethol Cymraeg Cyfoes!

Mae'r mis hwn yn nodi ail ben-blwydd y prosiect sy'n golygu ein bod yn dechrau cyfnod olaf y gwaith. Dros y 18 mis nesaf, byddwch yn dechrau gweld llawer mwy o weithgareddau gan y tîm wrth i rai darnau o'n gwaith gael eu rhyddhau. Mae nifer o sioeau teithiol ac arddangosiadau o'r data ac offer chwilio'r corpws ar y gweill hefyd – gweler ein gwefan, ein cyfryngau cymdeithasol a rhifynnau'r cylchlythyr am ragor o wybodaeth!



Mae'r rhifyn hwn o'r cylchlythyr yn rhoi'r diweddaraf am newyddion a datblygiadau o'r prosiect. Mae'n cynnwys manylion ein ymgais llwyddiannus i ddenu arian ar gyfer cynorthwywyr ymchwil israddedig; lansiad ein tagiwr rhannu ymadrodd; CyTag; a rhai diweddariadau o Becyn Gwaith 3 (WP3). Cewch hefyd gyfle i gwrdd ag aelod arall o'r tîm (Tom Cobb) ond, yn anffodus, mae'n rhaid inni hefyd ddweud hwyl fawr i un...

Mwynhewch y cylchlythyr!

Dr Dawn Knight

Ffarwel Dr Jeremy Evas

Ag yntau'n fis Mawrth, mae newidiadau ar droed ar dîm prosiect CorCenCC. Bydd Dr Jeremy Evans yn gadael y tîm ac yn ei le daw cyfaill a chydweithiwr iddo o Ysgol y Gymraeg, Dr Jonathan Morris. Bu'r ddau'n cydweithio'n ddiweddar ar Ymchwil i'r Amodau sydd yn Dylanwadu ar Arferion Teuluoedd mewn perthynas â Throsglwyddo a Defnyddio'r Gymraeg (Llywodraeth Cymru 2017). Ar hyn o bryd, mae Jeremy wedi'i secondio allan o'r Brifysgol fel Pennaeth Hybu'r Gymraeg yn Llywodraeth Cymru. Dywedodd Jeremy "Mae'n flin gen i adael prosiect mor ddiddorol ac rwy'n dymuno pob llwyddiant i chi wrth i chi gasglu data a'u dadansoddi. *Parsiwch* fy nymuniadau gorau i'r tîm

+ Newyddion a digwyddiadau

Croeso i'r tîm: Dr Jonathan Morris

Yn camu i mewn i esgidiau Jeremy, ar y 1af o Fawrth, 2018, ymunodd Dr Jonathan Morris â'r prosiect fel Cyd-Ymchwilydd (CI). Mae Jonathan yn Ddarlithydd y Coleg Cymraeg Cenedlaethol mewn Ieithyddiaeth ac Ieithyddiaeth Gymhwysol yn Ysgol y Gymraeg, Prifysgol Caerdydd. Mae Jonathan yn helpu cydlynu casglu data, felly bydd yn cyfrannu'n bennaf at Becyn Gwaith 1 (WP1). Gofynnom i Jonathan roi ychydig o eiriau ar ymuno â'r tîm a dyma oedd ganddo i'w ddweud:

Su' ma'i! Rwyf yn ddarlithydd y Coleg Cymraeg Cenedlaethol mewn ieithyddiaeth ac yn gweithio ers pum mlynedd yn Ysgol y Gymraeg, Prifysgol Caerdydd. Rwyf yn dod o Wrecsam yn wreiddiol ond cefais fy magu nid nepell o'r ffin yn Swydd Caer. Roeddwn yn hoffi ieithoedd yn yr ysgol ac fe symudais i Fanceinion er mwyn gwneud gradd mewn Ffrangeg ac Almaeneg. Fel rhan o'r radd, treuliais amser yn y Swistir ac, o ganlyniad, dechreuais ymddiddori go iawn mewn dwyieithrwydd. Cwblheais PhD ar amrywio seinegol a ffonolegol yn lleferydd siaradwyr dwyieithog Cymraeg-Saesneg yn 2013 a gweithiais yn rhan amser fel tiwtor Cymraeg i Oedolion.

Yn Ysgol y Gymraeg, rwyf yn dysgu ar fodiwlau megis 'Caffael Iaith' a 'Sosioieithyddiaeth' ac yn ymchwilio i'r ffactorau cymdeithasol (megis ardal a chefndir ieithyddol) sy'n dylanwadu ar sut y mae siaradwyr a dysgwyr y Gymraeg yn siarad yn eu dwy iaith. Mae gennyf ddiddordeb mewn cymdeithaseg iaith hefyd a bûm yn rhan o brosiect ar drosglwyddo'r Gymraeg ar yr aelwyd. Mae bod yn rhan o brosiect CorCenCC yn hynod gyffrous – mae'r prosiect yn cynnig cyfle unigryw i bobl o bob cefndir ac o bob cwr o Gymru gyfrannu at ddyfodol yr iaith.

Llwyddiannau Rhaglen Cyfleoedd Ymchwil i Israddedigion Caerdydd (CUROP)!

Rydym yn falch o gyhoeddi ein bod wedi derbyn cyllid gan Raglen Cyfleoedd Ymchwil Israddedig Caerdydd (CUROP), ar gyfer dau leoliad i fyfyrwyr dros yr haf. Mae CUROP yn galluogi myfyrwyr israddedig o Brifysgol Caerdydd i ennill profiad cyflogedig o ymchwil academaidd drwy wneud cais i gymryd rhan mewn prosiectau, o dan oruchwyliaeth. Enw'r prosiect cyntaf yw Corpws Cenedlaethol Cymraeg Cyfoes: National Corpus of Contemporary Welsh - tagio semantig ac anodi data. Bydd yr ymchwilydd CUROP ar y prosiect hwn yn cyfrannu at Becyn Gwaith 3 (WP3), yn gweithio ar adeiladu corpws o safon euraidd ac yn astudio materion semantig.

Enw'r ail brosiect yw Corpws Cenedlaethol Cymraeg Cyfoes: National Corpus of Contemporary Welsh - canolbwyntio ar ddata llafar. Bydd yr ymchwilydd CUROP ar y prosiect hwn yn cael ei oruchwylio gan arweinydd prosiect CorCenCC, Dr Dawn Knight, a Dr Lowri Williams, Cynorthwyydd Ymchwil sy'n rhan o'r tîm sy'n gyfrifol am lunio'r corpws 10 miliwn o eiriau. Fel y gwyddoch, ein nod yw casglu data gan gyfranwyr o wahanol oedrannau, lleoliadau daearyddol a galwedigaethau. Rôl yr ymchwilydd CUROP ar y prosiect hwn fydd cyfrannu tuag at y nod hwn. Bydd yr ymgeisydd llwyddiannus yn cynorthwyo gyda'r tasgau canlynol: (1) Casglu data, (2) Trawsgrifio a (3) Adfyfyrio ar (1) a (2) gan awgrymu mân newidiadau lle bo'n briodol.

Bydd ymchwilwyr CUROP ar y ddau brosiect hefyd yn helpu i hyrwyddo CorCenCC yn y gymuned ehangach, i annog cyfraniadau pellach i'r corpws ac i ysgogi diddordeb yn y defnydd o'r adnodd yn y dyfodol. Byddwn yn recriwtio myfyrwyr i'r prosiectau CUROP hyn yn ystod y pythefnos nesaf. Yn y cyfamser, os hoffech ddysgu mwy am y cynllun, gweler rhagor o wybodaeth yma: http://bit.ly/2pqN61A

Lleoliad SPIN (Interniaeth Abertawe â thâl): Croeso i Sioned!

Mae lleoliadau SPIN yn rhan o gynllun interniaeth Prifysgol Abertawe sy'n cael ei redeg gan SEA (Academi Cyflogadwyedd Abertawe). Mae lleoliadau'n para am bedair wythnos (neu'r oriau cyfatebol), maen nhw'n cynnig tâl ac mae modd gweithio'n rhan amser yn ystod y tymor neu'n amser llawn yn ystod y gwyliau/wythnosau heb addysgu. Rydym wrth ein boddau fod ein cais i SPIN ar gyfer lleoliad i CorCenCC wedi llwyddo ac wedi ein galluogi i hysbysebu am fyfyriwr i gyfrannu at y prosiect, gan ganolbwyntio ar weithio gyda thîm Abertawe ar (i) casglu data a (ii) cyfweliadau â defnyddwyr y corpws.

Roedd yn dda gennym dderbyn nifer o geisiadau cryf, a gallwn gyhoeddi, bellach, fod Sioned Johnson-Dowdeswell wedi'i phenodi ac fe fydd yn gweithio gyda Steve, Tess, Jenny a Mair ym Mhrifysgol Abertawe dros y misoedd nesaf yma.

Bydd Sioned yn gweithio gydag aelodau o dîm CorCenCC i recordio siaradwyr Cymraeg wrth iddyn nhw ddefnyddio'r iaith mewn sefyllfaoedd bob dydd ar draws Cymru (yn enwedig yn ei hardal enedigol Caerfyrddin). Hefyd, bydd hi'n helpu gyda'r gwaith trawsgrifio ac ymdrechion ehangach y prosiect i ymgysylltu â'r cyhoedd yn gyffredinol.

Yn ogystal â hyn, bydd Pecyn Gwaith 4 (y pecyn cymorth pedagogaidd) yn elwa o fewnbwn Sioned. Rydym yn rhagweld y bydd hi'n helpu o ran dadansoddi data o gyfweliadau gydag athrawon a thiwtoriaid, yn casglu eu barn i'n cynorthwyo i lywio datblygiad y deunyddiau pedagogaidd. Bydd y lleoliad hwn yn hybu'r cydweithio rhwng Adrannau Ieithyddiaeth Gymhwysol a'r Gymraeg yn Abertawe ac yn cryfhau'r cysylltiadau rhwng Pecynnau Gwaith 1 a 4 ar draws yr holl sefydliadau.

I gael mwy o wybodaeth am raglen SPIN, ewch i: https://tinyurl.com/yaotju43

Croeso i dîm CorCenCC, Sioned! Dyma air neu ddau am Sioned:

Fy enw i yw Sioned. Rwy'n fyfyrwraig flwyddyn gyntaf ym Mhrifysgol Abertawe ac yn astudio Cymraeg. Dw i'n dod o dref Caerfyrddin ac yn parhau i fyw yno wrth astudio yn y Coleg. Rwyf wedi bod yn ddigon ffodus i gael y cyfle i weithio ar brosiect CorCenCC ac rwy'n hynod o ddiolchgar am hynny. Er fy mod yn ymddiddori yn yr iaith Gymraeg, mae hefyd gen i ddiddordeb mewn anifeiliaid, mae gen i ddau gi a phum cath!!! Eddie yw enw'r gath fach yn y llun, wedi'i enwi ar ôl Ed Sheeran, am resymau amlwg!

+ Diweddariad cyffredinol

Y diweddaraf ar CySemTagger

Gan Scott Piao a Paul Rayson

Mae datblygiad y tagiwr semantig Cymraeg (a enwir yn CySemTagger) yn mynd yn ei flaen yn WP3, gyda'r cynnydd cyflym diweddar o greu data prawf o ansawdd uchel a fydd yn hwyluso'r datblygiad a'r gwerthusiad o ddulliau newydd. Mae Lowri a Jenny wedi gwneud gwaith ardderchog o anodi pob gair yn gywir yn y Corpws Safon Aur gyda chategori/au semantig cywir sy'n seiliedig ar y cynllun USAS. Mae adnodd o'r fath hon yn hanfodol er mwyn datblygu offeryn semantig a fydd yn ein galluogi i gymharu canlyniadau'r anodi awtomatig â barn a gwybodaeth arbenigwyr dynol. Y cam nesaf yw cymhwyso amryw o algorithmau cyd-destunol i'r data arbrofol hyn er mwyn inni benderfynu ar y dulliau gorau posibl ac er mwyn i'r tagiwr gyflawni'r safon uchaf posibl o ran cywirdeb.

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

Mae'r tabl ar y dde yn dangos sampl o destun Cymraeg sydd wedi'i anodi gyda chategorïau semantig. Mae'r drydedd a'r bedwaredd golofn yn cynnwys anodiadau a ddarperir gan y CySemTagger. Mae'r anodiadau hyn yn dynodi'r categori/au semanteg USAS perthnasol ar gyfer pob gair ac yn nodi a yw gair yn rhan o ymadroddion aml-eiriau (gan gynnwys idiomau a grwpiau o eiriau sy'n cyfleu un cysyniad semantig). Fel offeryn awtomatig, mae CySemTagger yn gwneud camgymeriadau, ac felly mae angen gwella'r canlyniadau drwy ddefnyddio data arbrofol o ansawdd uchel.

CyTag: y lansiad

Gan Steven Neale

Rydym yn lansio'r wefan ar gyfer CyTag y mis hwn, ein ffynhonnell agored o adnoddau tagio ar gyfer y Gymraeg. Mae CyTag wedi ei datblygu fel y brif dasg i'w chyflawni yn WP2. Ei nod yw creu tagiau rhannu ymadrodd (POS) a chyfres bwrpasol o dagiau fydd yn cael eu defnyddio i dagio CorCenCC. http://cytag.corcencc.org yw cyfeiriad y wefan, lle gallwch gael gwybod rhagor am y tagiau a rhoi cynnig ar ein henghraifft ar-lein!

Ar hyn o bryd, mae CyTag yn cynnwys rhannwr testun, adnodd sy'n hollti brawddegau, tokeniser, a thag y POS ei hun. Mae'r wefan yn seiliedig ar reolau ac mae'n gweithio drwy ddefnyddio gwybodaeth eirfaol o eiriadur GPL Eurfa i ddod o hyd i holl dagiau POS posibl ar gyfer symbol penodol. Mae wedyn yn mynd ati i ychwanegu dilyniannau o reolau at y symbolau sy'n penderfynu rhwng yr opsiynau hynny sy'n seiliedig ar y tagiau POS a/neu'r nodweddion geiriadurol o'r symbolau eraill o'i amgylch. Rydym wedi treialu CyTag gan ddefnyddio casgliad bach o frawddegau enghreifftiol, ac rydym yn fodlon iawn ar ei chywirdeb - sydd ar hyn o bryd dros 95%! Yn fwy cyffrous byth, cafodd papur yn disgrifio'r CyTag ei dderbyn i'w gyhoeddi yng nghynhadledd Adnoddau Iaith a Gwerthuso (LREC) 2018 a gynhelir yn Miyazaki, Japan ym mis Mai. Mae'r papur - a ysgrifennwyd gan Steven Neale, Kevin Donnelly, Gareth Watkins a Dawn Knight - yn cyflwyno trosolwg dechnegol fanwl o CyTag yn ogystal â gwerthusiad manwl o gywirdeb y gwaith. Rydym yn edrych ymlaen at ei gyflwyno i'r gymuned ehangach, ac at gael adborth defnyddiol er mwyn ei ddatblygu ymhellach!

+ Cwrdd â'r tîm: Tom Cobb, Prifysgol Quebec yn Montreal, Ymgynghorydd Prosiect

Mae'n eironig fy mod yn ymwneud â phrosiect Corpws Cenedlaethol y Gymraeg oherwydd er fy mod yn dod o Ganada, dechreuodd fy ngyrfa addysgol gyfan yng Nghymru - Coleg Prifysgol Gogledd Cymru ym Mangor, lle astudiais am gymhwyster TAR mewn addysgu Saesneg tua 1978. Roedd y cwrs yn cynnwys mynd allan ar sawl interniaeth addysgu mewn ysgolion, ac roedd gweinyddwyr fy ngholeg bob amser yn chwilio am ysgolion ble byddai hyfforddai heb unrhyw wybodaeth o'r Gymraeg o gwbl (fi!) yn dderbyniol. Roedd y rhain fel arfer o fewn ychydig filltiroedd i'r ffin â Lloegr, fel arfer yng Nghlwyd. Dros y cyfnod hwnnw, fodd bynnag, fe ddysgais ychydig o Gymraeg, byw mewn cartref rhannol Gymraeg a mynychu tafarn oedd i bob pwrpas yn uniaith Gymraeg.

Fodd bynnag, yn fuan iawn ar ôl hyfforddi i fod yn athro Saesneg mewn ysgol uwchradd dychwelais i Ganada i ddod yn

ddarlithydd prifysgol sesiynol mewn cyfansoddi Saesneg, swydd nad oeddwn, wrth gwrs, wedi derbyn unrhyw hyfforddiant ar ei chyfer. Ond roeddwn i'n gallu gweld bod gan hyfforddiant ysgrifennu fwlch mawr a gwag ar gyfer dysgu gyda chymorth cyfrifiadur (CALL), er enghraifft i ymarfer gweithrediadau lled-fecanyddol fel gosod atalnodi a nodi darnau o frawddegau. Ond doeddwn i ddim yn rhaglennydd cyfrififadurol ar y pryd, felly parhaodd fy myfyrdodau ar swyddogaeth cyfrifiaduron yn ... ddamcaniaethol.

Y prif beth arall a ddysgais yn y gwaith hwn mewn prifysgol yng Nghanada oedd nad oedd yn talu'n dda iawn, felly yn fuan iawn gadewais i fynd i wledydd fel Saudi Arabia ac Oman lle'r oedd y cyflogau'n well (a'r amodau'n waeth, ond stori arall yw honno). Tua'r adeg hon, prynais fy nghyfrifiadur cyntaf, ac yn y prynhawniau poeth eang yn Saudi pan nad oedd neb yn gweithio, dechreuais ddysgu sut i raglennu cyfrifiadur. Gwnes rai pethau syml i ddysgwyr, gan gludo cyfrifiadur Osborne maint siwtces 'llusgadwy' (nid cludadwy) ar draws yr anialwch rhwng y coleg a'r maes parcio er mwyn i fyfyrwyr gael profi llawenydd CALL ar hap. Erbyn diwedd y profiad hwn ymddangosodd y Macintosh SE30 gyda Hypercard wedi'i lwytho ac yn fuan roeddwn i'n creu gemau dysgu a llawer o feddalwedd arall ac yn trefnu i ambell fyfyriwr chwarae gyda'r rhain am ychydig er mwyn i fi allu gweld mannau gwan y rhyngwynebau.

Ychydig yn ddiweddarach, mewn ymgais i aros yn y Gwlff ond gadael Saudi Arabia, symudais draw i anialwch dymunol Swltaniaeth Oman, i swydd debyg ond gyda labordy Macintosh i'w sefydlu a'i redeg mewn coleg newydd. Roedd hwn yn amseru perffaith ar gyfer fy egin sgiliau codio, a hefyd roedd Canolfan Iaith Prifysgol Sultan Qaboos yn ferw o waith corpws a mynegeirio (roedd John Sinclair wedi ymweld â'r lle hyd yn oed, ond cyn fy amser i). Ymhen ychydig roeddwn i wedi dechrau ymgorffori cysyniadau corpws yn fy ngemau a'm hymarferion, prosiect a arweiniodd at PhD mewn technoleg addysgol ac at lawer o flynyddoedd dedwydd yn ceisio meddwl am ffyrdd i ennyn diddordeb dysgwyr mewn 'data mawr' wrth iddynt ganfod datrysiadau i'w cwestiynau am iaith.

Arweiniodd y gwaith corpws a 'CALL' yma at swydd go iawn yn ôl yng Nghanada ac mae wedi arwain, bellach, at rôl ar ochr diwtorial prosiect y Corpws Cymraeg. Unwaith mae corpws gennych, beth am ei ddefnyddio fel offeryn dysgu? Ymhlith pethau eraill, wrth gwrs, ar gyfer gyrfa mewn ieithoedd a ddechreuodd yng Nghymru, does dim modd cael gwell cymesuredd!

+ 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 a Jonathan Morris; y Cynorthwywyr Ymchwil - Steven Neale, Jennifer Needs, Mair Rees, Scott Piao a Lowri Williams; y myfyriwr 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