Y Corpws Cenedlaethol Cymraeg Cyfoes

Adroddiad y Prosiect, Hydref 2020

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

1. Cyflwyniad

1.1. Diben yr adroddiad hwn

Bydd yr adroddiad hwn yn rhoi trosolwg o brosiect Corpws Cenedlaethol Cymraeg Cyfoes (CorCenCC) a'r adnodd corpws ar-lein a ddatblygwyd o ganlyniad i waith y prosiect. Bydd yr adroddiad yn amlinellu sylfeini damcaniaethol yr ymchwil, gan ddangos sut mae'r prosiect wedi adeiladu ar y theori hon a'i hymestyn. Byddwn hefyd yn codi ac yn trafod rhai o'r cwestiynau gweithredol allweddol a gododd wrth i'r prosiect fynd rhagddo, gan amlinellu'r ffyrdd y cawsant eu hateb, effaith y penderfyniadau hyn ar yr adnodd sydd wedi'i lunio, a'r cyfraniad tymor hwy y byddant yn ei wneud i arferion wrth adeiladu corpws. Yn olaf, byddwn yn trafod rhai o gymwysiadau'r gwaith a'r modd o'i ddefnyddio, gan amlinellu'r effaith y mae'r CorCenCC yn debygol o'i chael ar amrediad o unigolion a grwpiau defnyddwyr gwahanol.

1.2. Trwydded

Trwyddedir corpws CorCenCC a'r offerynnau meddalwedd cysylltiedig dan Creative Commons CC-BY-SA f4 ac felly maent yn rhydd ar gyfer eu defnyddio gan gymunedau proffesiynol ac unigolion sydd â diddordeb mewn iaith. Darperir cymwysiadau a chyfarwyddiadau pwrpasol ar gyfer pob offeryn (cyfeiriwch at adran 10 yr adroddiad hwn am ddolenni i'r holl offerynnau). Wrth adrodd gwybodaeth a ddeilliodd o ddefnyddio data a/neu offerynnau corpws CorCenCC, dylid cydnabod CorCenCC yn briodol (gweler 1.3.).

- I gael mynediad i'r corpws, ewch i: www.corcencc.cymru/archwilio/
- I gael mynediad i wefan GitHub: https://github.com/CorCenCC
    - Mae GitHub yn wasanaeth sydd yn y cwmwl sy'n galluogi datblygwyr i storio, rhannu a rheoli eu cod a'u setiau data.

1.3. Cyfeirio at CorCenCC

Rhaid rhoi cydnabyddiaeth briodol wrth ddefnyddio data a/neu offerynnau corpws CorCenCC. Defnyddiwch y canlynol wrth gyfeirio at gorpws CorCenCC a'r adroddiad prosiect presennol:

- Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., Thomas, E-M., Lovell, A., Morris, J., Evas, J., Stonelake, M., Arman, L., Davies, J., Ezeani, I., Neale, S., Needs, J., Piao, S., Rees, M., Watkins, G., Williams, L., Muralidaran, V., Tovey, B., Anthony, L., Cobb, T., Deuchar, M., Donnelly, K., McCarthy, M. a Scannell, K.

(2020). CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes – *The National Corpus of Contemporary Welsh*. Prifysgol Cardiff. http://doi.org/10.17035/d.2020.0119878310

- **Adroddiad:** Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I. a Thomas, E. M. (2020). The National Corpus of Contemporary Welsh: Project Report | Y Corpws Cenedlaethol Cymraeg Cyfoes: Adroddiad y Prosiect. *arXiv:2010.05542*, Hydref 2020.

Gellir dod o hyd i gyhoeddiadau eraill yn adran 10 yr adroddiad hwn ac ar dab 'Allbynnau' gwefan CorCenCC: www.corcencc.cymru/allbynnau/

1.4. Cydnabyddiaethau

Cafodd yr ymchwil y seiliwyd yr adroddiad hwn arni, a'r adnodd corpws ar-lein sy'n cyd-fynd â hi, eu hariannu gan Gyngor Ymchwil Economaidd a Chymdeithasol y Deyrnas Unedig (ESRC) a'r Cyngor Ymchwil i'r Celfyddydau a'r Dyniaethau (AHRC) fel *y Corpws Cenedlaethol Cymraeg Cyfoes: Dull cymunedol o adeiladu corpws ieithyddol* (Rhif Grant ES/M011348/1). Gellir dod o hyd i wybodaeth am aelodau tîm y prosiect yn www.corcencc.cymru/cysylltiadau. Ni fyddai prosiect CorCenCC wedi bod yn bosibl heb eu mewnbwn, eu harbenigedd, eu brwdfrydedd a'u colegoldeb.

Rydym hefyd am gydnabod Prifysgol Caerdydd a Phrifysgol Abertawe am eu cyfraniad i ysgoloriaethau doethuriaeth, gan ein galluogi i gynnwys ymchwilwyr ôl-raddedig yn nhîm y prosiect. Rydym yn diolch yn arbennig ac yn bersonol i'r cydweithwyr hynny yn ein holl brifysgolion perthnasol sydd wedi rhoi'n hael o'u hamser a'u cymorth i ni yn ystod camau hanfodol bwysig y prosiect.

Rhaid rhoi'r diolch am gyflawni prosiect CorCenCC hefyd i'n rhanddeiliaid prosiect (gweler 2.2.), ac yn enwedig y bobl hynny o grŵp cynghori'r prosiect: sydd nid yn unig wedi bod yn hael wrth hwyluso'r gwaith o gasglu data, neu ei gyfrannu'n uniongyrchol, ond sydd hefyd wedi bod yn galonogol yn eu cefnogaeth o safbwynt nodau'r prosiect a'u hymgysylltiad â'r broses gynllunio, yn ogystal â'u hymroddiad i gynaliadwyedd a pharhad CorCenCC.

2. Gweledigaeth ac amcanion

2.1 Trosolwg o'r prosiect

Prosiect rhyngddisgyblaethol ac amlsefydliadol yw CorCenCC sydd wedi creu corpws ffynhonnell agored o Gymraeg cyfoes ar raddfa eang. Yn y cyd-destun hwn, corpws yw casgliad o enghreifftiau o iaith lafar, ysgrifenedig a/neu e-iaith o gyd-destunau bywyd go iawn sy'n galluogi defnyddwyr i nodi ac archwilio iaith fel y mae'n cael ei defnyddio mewn gwirionedd yn hytrach na dibynnu ar reddf neu gyfarwyddiadau rhagnodol o ran sut y 'dylai' gael ei defnyddio. Mae corpysau'n ein galluogi i ymchwilio i sut yr ydym yn defnyddio iaith ar draws sawl genre a chyfrwng cyfathrebu gwahanol (h.y. ar lafar, yn ysgrifenedig neu'n ddigidol), a sut mae'n amrywio yn ôl y siaradwr/awdur a diben y cyfathrebu. Defnyddir y dull sy'n seiliedig ar dystiolaeth hwn gan ymchwilwyr academaidd, geiriadurwyr, athrawon,

dysgwyr iaith, aseswyr, datblygwyr adnoddau, llunwyr polisïau, cyhoeddwyr, cyfieithwyr ac eraill, ac mae'n hanfodol i waith datblygu technolegau fel llunio testun rhagfynegol, offerynnau prosesu geiriau, cyfieithu peirianyddol, adnabod llais ac offerynnau chwilio ar y we.

Cyn adeiladu CorCenCC, roedd nifer o gorpysau'r Gymraeg yn bodoli, gan gynnwys corpws Siarad (Deuchar ac eraill, 2018), a oedd yn cynnwys 460,000 o eiriau llafar, Corpws Cymraeg Crúbadán (Scannell, 2007), a oedd yn seiliedig ar e-iaith ac yn cynnwys 24 miliwn o eiriau, a Chorpws Llafar Paldaruo (Cooper ac eraill, 2019), a gafodd ei gyfrannu'n dorfol ac a oedd yn cynnwys darnau o destun a oedd wedi'u darllen yn uchel. Rhoddwyd ystyriaeth i p'un a ellid integreiddio'r rhain yn amcanion CorCenCC, neu eu halinio â nhw. Gweithiodd Deuchar a Scannell fel ymgynghorwyr ar gyfer y prosiect, a chafodd Canolfan Bedwyr (Bangor) ei chynrychioli yng ngrŵp cynghori'r prosiect. Fodd bynnag, o gofio y cafodd y corpysau blaenorol eu casglu er mwyn gwireddu nodau a gweledigaethau gwahanol, penodol a phwrpasol, ystyriwyd bod angen creu set ddata newydd a chyflawn.

CorCenCC yw corpws cyntaf y Gymraeg sy'n cwmpasu pob un o dair elfen iaith y Gymraeg cyfoes: llafar, ysgrifenedig ac iaith sydd wedi'i chyfryngu'n electronig (e-iaith). Mae'n cynnig ciplun o'r Gymraeg ar draws amrediad o gyd-destunau lle caiff ei defnyddio, e.e. sgyrsiau preifat, cymdeithasu fel grŵp, busnes a sefyllfaoedd gwaith eraill, ym maes addysg, yng nghyfryngau amrywiol y wasg, ac mewn mannau cyhoeddus. Mae'n cynnwys enghreifftiau o benawdau newyddion, gohebiaeth a negeseuon e-bost personol a phroffesiynol, gwaith ysgrifenedig academaidd, iaith lafar ffurfiol ac anffurfiol, postiadau ar flogiau a negeseuon testun (caiff cynnwys penodol y corpws ei drafod yn adran 3.3). Cafodd data am iaith ei samplu gan amrediad o siaradwyr a defnyddwyr Cymraeg, o bob rhanbarth o Gymru, o bob oedran a rhyw, o amrediad eang o swyddi ac o amrywiaeth o gefndiroedd ieithyddol (e.e. sut y daethant i siarad Cymraeg), i adlewyrchu amrywiaeth y mathau o destun a'r siaradwyr Cymraeg a geir yn y Gymru gyfoes. Yn y modd hwn, mae corpws CorCenCC yn darparu ffordd o rymuso defnyddwyr Cymraeg i ddeall yr iaith yn well, ac arsylwi arni, ar draws lleoliadau amrywiol, ac mae'n creu sylfaen dystiolaeth gadarn ar gyfer addysgu Cymraeg cyfoes i'r bobl hynny sydd am ei defnyddio. Dros amser, gallai'r corpws gyfrannau'n sylweddol at drawsnewid y Gymraeg fel yr iaith drafod ym meysydd y cyhoedd, masnach, addysg a llywodraeth.

I'r perwyl hwnnw, nod CorCenCC yw galluogi defnyddwyr cymunedol, er enghraifft, i ymchwilio i amrywiadau neu fympwyon tafodieithol yn eu hiaith eu hunain; defnyddwyr proffesiynol i broffilio testunau i weld pa mor ddarllenadwy ydynt neu ddatblygu offerynnau iaith digidol; dysgwyr yr iaith i ddysgu o fodelau go iawn o Gymraeg; ac ymchwilwyr i ymchwilio i batrymau o ran sut caiff iaith ei defnyddio a'i newid. Rhagwelir y bydd y corpws hefyd yn datgelu mewnwelediadau newydd i batrymau geirfa ac iaith y Gymraeg ac y bydd yn adnodd pwysig ar gyfer addysgu'r Gymraeg i siaradwyr y mae'n iaith gyntaf ganddynt a siaradwyr Cymraeg newydd. Mae'r effaith amlochrog bosibl hon wedi bod yn bosibl oherwydd cyfraniad sylweddol CorCenCC ar y lefel fethodolegol, wrth ymestyn cwmpas, perthnasedd a seilwaith dylunio corpysau ieithyddol. Yn benodol, mae'r prosiect wedi golygu datblygu offerynnau a phrosesau newydd pwysig, gan gynnwys gwaith cynllunio corpws unigryw a yrrir gan ddefnyddwyr lle cafodd data am iaith ei gasglu a'i ddilysu'n dorfol, a phecyn cymorth addysgeg (Y Tiwtiadur), yr oedd wedi'i ymgorffori, a ddatblygwyd

drwy ymgynghori â chynrychiolwyr o bob grŵp o ddefnyddwyr academaidd a chymunedol a ragwelwyd (gweler Knight ac eraill, 2020 – ceir manylion yn adran 10.2. isod – am drafodaeth fanwl o gynllun CorCenCC a yrrir gan ddefnyddwyr).

2.2. Tîm y prosiect

Roedd y canlynol ynghlwm wrth brosiect CorCenCC: pedwar sefydliad academaidd (sef Prifysgol Caerdydd, Prifysgol Abertawe, Prifysgol Caerhirfryn a Phrifysgol Bangor), un prif ymchwilydd (Dawn Knight), dau gyd-ymchwilydd (Tess Fitzpatrick a Steve Morris) a ffurfiodd, ynghyd â'r prif ymchwilydd, Dîm Rheoli CorCenCC, a chyfanswm o saith cyd-ymchwilydd eraill (Irena Spasić, Paul Rayson, Enlli Môn Thomas, Alex Lovell, Jonathan Morris, Jeremy Evas a Mark Stonelake), deg cynorthwyydd/cysylltai ymchwil, a 180+ o drawsgrifwyr a fu'n gweithio yn ystod y prosiect.

Yn ogystal, roedd chwe ymgynghorydd, dau fyfyriwr doethuriaeth, pedwar myfyriwr israddedig ar leoliad haf, pedwar aelod o staff cymorth gwasanaethau proffesiynol, a dau wirfoddolwr prosiect. Bu'r prosiect ar ei ennill hefyd gan gyfraniadau a chymorth gan gynrychiolwyr o amrediad o randdeiliaid, gan gynnwys Llywodraeth Cymru, Cynulliad Cenedlaethol Cymru, y BBC, S4C, CBAC, Cymraeg i Oedolion, Gwasg y Lolfa, SaySomethinginWelsh a Geiriadur Prifysgol Cymru, drwy Grŵp Cynghori Prosiect. Nia Parry (cyflwynydd, cynhyrchydd ac ymchwilydd teledu; 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 (bardd; Athro Llenyddiaeth Cymraeg a Saesneg; cyn-Gadeirydd Llenyddiaeth Cymru) yw llysgenhadon swyddogol prosiect CorCenCC. Gellir gweld rhestr lawn o'r holl unigolion a fu ynghlwm wrth y prosiect yn www.corcencc.cymru/cysylltiadau/ – a chyfeirir at lawer ohonynt drwy gydol yr adroddiad hwn fel y bo'n berthnasol.

Cafodd y prosiect ei hwyluso gan dîm traws-sefydliadol cadarn, a gefnogodd waith recriwtio staff, rheoli ariannol, technoleg gwybodaeth (gan gynnwys cyfarpar, meddalwedd, cynnal a chadw'r gweinyddion a gwefannau), allgymorth y cyfryngau a chyfathrebu (gan gynnwys cydlynu datganiadau i'r wasg, a siarad ar y radio a'r teledu), arweiniad cyfreithiol o safbwynt ffurflenni, contractau a thrwyddedau, a chyfieithwyr a chyfieithwyr ar y pryd Cymraeg (a ddarparodd gyfieithiadau ysgrifenedig ar gyfer adroddiadau, dogfennau prosiect allweddol ac allbynnau eraill, a chyfieithu ar y pryd yn ystod holl gyfarfodydd y Tîm Prosiect Cyfan a digwyddiadau rhannu gwybodaeth â'r cyhoedd). Cafodd dros 210 o adroddiadau eu hysgrifennu, ddwywaith yr wythnos, yn ystod y prosiect, a oedd yn nodi manylion y gwaith a gyflawnwyd, problemau a risgiau allweddol, y gwaith i'w wneud, syniadau newydd, meddyliau a chyfleoedd; cynhaliwyd deg cyfarfod Tîm Prosiect Cyfan, a digwyddodd ymhell dros 100 o gyfarfodydd ychwanegol. Cafodd saith rhestr bostio fewnol eu llunio er mwyn hwyluso'r cyfathrebu rhwng aelodau'r tîm a oedd wedi'u lleoli ar safleoedd gwahanol, ynghyd â siart Gantt ganolig ar gyfer y prosiect cyfan, er mwyn cofnodi ac olrhain y pethau yr oedd angen eu cyflawni, a cherrig milltir prosiect allweddol, ar y cyd. Er mwyn cynnal y cyfathrebu â'r cyhoedd a rhanddeiliaid eraill, cafodd 24 rhifyn o newyddlen brosiect eu cyhoeddi a'u dosbarthu i unigolion, a'u lanlwytho i'r brif wefan. Roedd y newyddlenni'n darparu darllenwyr â'r wybodaeth ddiweddaraf am waith casglu data, yn adrodd ar gyflwyniadau a phrif gyflwyniadau a roddwyd (sef cyfanswm o 54 ohonynt), ac yn eu

cyflwyno i aelodau unigol y tîm drwy ein pwt 'dyma'r tîm' rheolaidd. Diben hwn oedd cynnal diddordeb ac ymdeimlad o fuddsoddi yn y prosiect (yn unol â'r cynllun a yrrir gan ddefnyddwyr). Gwnaeth gwefannau'r prosiect (sef www.corcencc.cymru | www.corcencc.org), a ffrydiau Facebook a Twitter hefyd hwyluso â'r broses o ennyn diddordeb y cyhoedd, gan arwain at dros 140,000 o ymweliadau â'r gwefannau a 1,029 o ddilynwyr ar Twitter a 374 ar Facebook (hyd at fis Awst 2020). Er y cânt eu crybwyll yn olaf yma, aelodau pwysicaf y tîm CorCenCC estynedig yw'r 2,000+ o unigolion a gyfrannodd at y corpws.

2.3. Pecynnau gwaith (WP)

Cafodd y gwaith ar brosiect CorCenCC ei ddosbarthu ar draws chwe phecyn gwaith (WP) cydgysylltiedig, yr oedd gan bob un ohonynt dasgau, nodau ac amcanion penodol. Dan arweiniad Knight, roedd WP0 yn cynnwys gweithgareddau dylunio, rhychwantu a hyfforddi parhaus, ac yn cynnwys holl aelodau tîm y prosiect. Roedd cynnwys y pecynnau gwaith eraill fel a ganlyn:

- WP1: Casglu, trawsgrifio ac anonymeiddio'r data
- WP2: Datblygu'r set dagiau / tagiwr rhannau ymadrodd
- WP3: Datblygu tagiwr semantig ar gyfer y Gymraeg a thagio'r holl ddata yn semantig
- WP4: Cwmpasu, dylunio ac adeiladu Y Tiwtiadur
- WP5: Adeiladu'r seilwaith i letya CorCenCC a chreu'r corpws

Er y cafodd y gwaith ei ddosbarthu ar draws y pecynnau gwaith hyn, roedd cydweithwyr yn meddu ar gyd-ddealltwriaeth o'r weledigaeth a rennir ar gyfer y prosiect, gan weithio ar y cyd er mwyn ei chyflawni, ac roedd lefel sylweddol o gyd-ddibyniaeth rhwng y pecynnau gwaith yr oedd yn gofyn am waith trafod a chydlynu. Er enghraifft, roedd WP3 yn adeiladu ar y gwaith ymchwil a wnaethpwyd yn WP1 o safbwynt casglu data ar gyfer y corpws, a defnyddiodd dagiwr rhannau ymadrodd WP2 fel y cam cyntaf ym mhroses dadansoddi'r data am y Gymraeg yn semantig. Yna, llywiodd allbwn WP3 waith WP4 ar gyfer y pecyn cymorth pedagogaidd ar-lein (Y Tiwtiadur), a ddefnyddiodd lefelau lluosog o anodiadau'r corpws i wella ymgysylltiad athrawon a dysgwyr â'r pecyn cymorth, a'r hyn y gellid ei ddefnyddio ar ei gyfer. Yn ogystal, llywiodd y corpws yr oedd wedi'i dagio'n semantig yn WP3 seilwaith y corpws a ddatblygwyd yn WP5. Yn yr adrannau canlynol, byddwn yn rhoi disgrifiad manwl o'r pecynnau gwaith, yn amlinellu eu prif nodau ac amcanion, ac yn myfyrio ar eu prif gyflawniadau, eu prif gyfraniadau a ffyrdd posibl o'u defnyddio. Ysgrifennwyd y disgrifiadau hyn gan arweinwyr perthnasol y pecynnau gwaith.

3. Pecyn Gwaith 1: Casglu, trawsgrifio ac anonymeiddio'r data

3.1. WP1: Disgrifiad

Prif waith WP1 oedd cyrchu, casglu a phrosesu'r data a oedd i'w gynnwys yn CorCenCC. Elfennau craidd y weithdrefn hon oedd i) creu fframwaith samplu'r prosiect; ii) sefydlu confensiynau trawsgrifio; iii) sicrhau dull cyson o safbwynt cydymffurfiaeth foesegol o ran casglu'r data. Cafodd WP1 ei arwain ar y cyd gan Morris a Knight, ac ymunodd tîm o

ymchwilwyr a oedd yn siarad Cymraeg â nhw, gan gynnwys y cyd-ymchwilwyr Evas, J. Morris, a Lovell, a'r cynorthwywyr/cysyllteion ymchwil Needs, Rees, Arman, Watkins a Williams (ar bwyntiau gwahanol yn ystod y prosiect). Darparodd Deuchar a McCarthy, sy'n arwain ym maes ieithyddiaeth corpws, gyngor ymgynghorol parhaus ar gyfer y cam hwn drwy gydol y prosiect a chafodd cymorth ychwanegol ei ddarparu gan sawl gwirfoddolwr ac intern prosiect.

3.2. WP1: Amcanion

Roedd nodau ac amcanion WP1 fel a ganlyn:

- dylunio fframwaith samplu ar gyfer y corpws
- cyrchu a chasglu data priodol
- cynllunio a chymhwyso protocolau trawsgrifio i ddata llafar

Y fframwaith samplu

Roedd yr Achos am Gefnogaeth yng nghynnig y prosiect yn cynnwys canllaw amlinellol i'n hamcanion mewn perthynas ag elfennau (llafar, ysgrifenedig ac e-iaith), genres a thestunau ieithyddol, ac amcangyfrif o faint o eiriau y byddai'n cael eu casglu o dan bob pennawd. Un o dasgau cyntaf WP1 oedd mireinio a datblygu'r canllaw ar-lein hwn ymhellach. Cafodd fframwaith samplu ei llunio er mwyn tanategu'r gwaith casglu data ar gyfer y prosiect, er mwyn sicrhau y byddem yn casglu amrediad o siaradwyr gwahanol ar draws cyd-destunau trafod a lleoliadau daearyddol gwahanol. Cynlluniwyd y fframwaith samplu i adlewyrchu demograffeg bresennol siaradwyr Cymraeg gan ddefnyddio'r wybodaeth cyfrifiad ddiweddaraf (ONS, 2011). Un o'r agweddau arloesol ar fframwaith samplu CorCenCC yw'r ystyriaeth fanwl y mae'n ei rhoi i'r meysydd lle defnyddir y Gymraeg. Mewn cyd-destun lle bo'r mwyafrif helaeth o siaradwyr Cymraeg yn ddwyieithog a bod gwasgariad daearyddol anghyson yn nhermau dwysedd siaradwyr, oedran siaradwyr a meysydd ieithyddol, roedd angen i'r fframwaith samplu adlewyrchu sefyllfa sosioieithyddol gyfredol yr iaith yn y modd mwyaf cywir posibl.

Cyrchu'r data

Cafodd y targedau ar gyfer y data llafar, ysgrifenedig ac e-iaith a oedd i'w gasglu ar gyfer CorCenCC, a'i ffynonellau, eu gyrru gan ein fframwaith samplu a chawsant eu llunio gan ymchwiliad cychwynnol i ble y siaredir y Gymraeg a lle ceir y defnydd mwyaf ohoni yn nhermau deunydd ysgrifenedig ac e-iaith. Mewn cyd-destun dwyieithog, gallai meysydd penodol fod wedi'u tangynrychioli (e.e. papurau newydd dyddiol cenedlaethol). Felly, roedd angen sicrhau bod y data'n adlewyrchiad gwirioneddol o'r hyn sydd ar gael i ddefnyddwyr yr iaith ac y mae modd ei gyrchu ganddynt, yn hytrach nag atgynhyrchu fframweithiau sydd â'r nod o greu corpysau mewn ieithoedd lle bo'r mwyafrif o siaradwyr yn unieithog.

Trawsgrifio

Roedd dau gam paratoadol: (i) creu confensiynau trawsgrifio ar gyfer y Gymraeg a (ii) recriwtio trawsgrifwyr (gweler 3.3. isod hefyd). Roedd heriau penodol ynghlwm wrth cam (i) am sawl rheswm:

- Yn yr iaith ysgrifenedig, mae awduron yn aml yn dynodi amrywiaeth ym mathau gwahanol o iaith lafar, fel bod yr un ystyr ieithyddol yn cael ei ysgrifennu mewn llawer o ffyrdd gwahanol. Enghraifft o hyn byddai amser presennol y person cyntaf unigol ar gyfer 'bod'. Yn Saesneg, gellid cyfleu hyn ar ffurf '*I am*' neu '*I'm*'. Byddai Cymraeg ffurfiol ysgrifenedig yn rhoi '*Yr wyf (i)*' neu '*Rwyf i*'. Fodd bynnag, ceir y posibiliadau canlynol ar gyfer Cymraeg llafar: '*Rydw i / Dw i / Rwy / Wy / Fi*'. Byddai awduron yn ysgrifennu'r ffurfiau hyn i gynrychioli siaradwyr o ardaloedd gwahanol a gellir arsylwi arnynt mewn llenyddiaeth a chyfryngau ysgrifenedig eraill. Roedd angen i'r confensiynau trawsgrifio felly allu adlewyrchu'r realaeth hon wrth ei mynegeio at yr un ystyr, fel y gallai chwiliadau ddod o hyd i'r sylweddiadau gwahanol.
- O gofio bod CorCenCC yn cynnwys iaith electronig yn ogystal â data ysgrifenedig a llafar, a'r ffaith nad yw'r confensiynau ar gyfer cynrychioli iaith yn y cyd-destun electronig wedi'u sefydlu'n llawn mewn unrhyw iaith, roedd angen darparu ar gyfer yr achosion hyn mewn modd tebyg.
- Mewn perthynas â thrawsgrifio deunydd llafar, yr egwyddor gyffredinol a fabwysiadwyd oedd alinio'r hyn y cafodd ei glywed â'r sylweddiad ysgrifenedig agosaf o set a oedd yn dechrau â'r amrediad presennol o ffurfiau ysgrifenedig ond yr ategwyd ati yn ôl yr angen er mwyn sicrhau bod pob ffurf lafar wedi'i chipio'n briodol. Cafodd yr egwyddor hon ei mireinio a'i datblygu drwy sawl iteriad o gonfensiynau trawsgrifio CorCenCC.
- Ar ôl mabwysiadu'r egwyddor drawsgrifio gyffredinol hon, roedd angen sicrhau cysondeb ar draws tîm trawsgrifio CorCenCC. Roedd yn rhaid ymwrthod yn gadarn ag unrhyw duedd tuag at ragysgrifiadaeth neu nodi ffurf ysgrifenedig ffurfiol y Gymraeg.

Cafodd trawsgrifwyr eu recriwtio drwy ymgyrchoedd a oedd wedi'u targedu'n benodol at aelodau o'r proffesiwn cyfieithu (drwy newyddlen eu cymdeithas), myfyrwyr prifysgol, a'r bobl hynny a oedd wedi bod ynghlwm wrth y gwaith o drawsgrifio ar gyfer prosiectau eraill. Roedd yn rhaid i bob trawsgrifydd lwyddo mewn darn prawf rhagarweiniol (gan lynu wrth gonfensiynau trawsgrifio CorCenCC), a chafodd ansawdd ei sicrhau drwy wirio, ar hap, 25 y cant o'r holl waith a drawsgrifiwyd, gan wneud cywiriadau lle nodwyd problemau.

3.3. WP1: Cyflawniadau

Y fframwaith samplu

Mae Tabl 1 yn cynrychioli'r fframwaith samplu cychwynnol a gynigiwyd yn yr Achos am Gefnogaeth yng nghynnig y prosiect. Roedd y dosbarthiadau cychwynnol hyn yn seiliedig ar y dymuniad o ganolbwyntio ar gynrychioli'r iaith a siaredir yn ddifyfyr (ar lafar ac e-iaith) o'i chymharu ag iaith sydd wedi'i pharatoi (ysgrifenedig). Ar ôl cychwyn y prosiect, cafodd fframiau samplu mwy manwl eu datblygu ar gyfer cydrannau llafar (Tabl 2), ysgrifenedig (Tabl 3) ac e-iaith (Tabl 4) y corpws. Cafodd proses mireinio iteraidd y fframwaith samplu ei llywio gan grwpiau thematig a chategoreiddio trafod corpysau pwysig a oedd eisoes yn bodoli, gan gynnwys Corpws Cenedlaethol Prydain (BNC) 1994, Corpws Llafar Cenedlaethol Prydain 2014, CANCODE a CANELC (gweler Aston a Burnard, 1997,

McEnery ac eraill, 2017, Carter and McCarthy, 2004 a Knight et al., 2013). Darparodd fframweithiau'r corpysau hyn a oedd eisoes yn bodoli fan cychwyn defnyddiol ar gyfer archwilio i ba raddau y mae'n bosib defnyddio'r un grwpiau a chategorïau yng nghyd-destun ieithoedd sydd wedi'u lleiafrifo. Ceir gwybodaeth sydd â rhagor o fanylder am bob un o'r is-genres, a'r cyfiawnhad dros y dosbarthiadau arfaethedig hyn, yn Knight ac eraill, 2021a (gweler adran 10.2.).

Tabl 1. Fframwaith samplu cychwynnol ar gyfer CorCenCC (sydd wedi'i dynnu o'r Achos am Gefnogaeth).

Math	Ffynonellau enghreifftiol (amcangyfrif)	Geiriau	Cyfanswm
Llafar	Iaith drafod dysgwyr Cymraeg	600,000	4 miliwn
	Sgyrsiau â ffrindiau; â'r teulu; cyfweliadau a ddarlledwyd ar y teledu a sioeau siarad teledu (BBC); Cymraeg yn y gweithle	400,000 yr un	
	Sioeau radio'r BBC; cyfarfyddiadau gwasanaethau	400,000 yr un	
	Galwadau ffôn; rhyngweithiadau mewn dosbarthiadau ysgolion cynradd ac uwchradd a cholegau trydyddol ac mewn dosbarthiadau i oedolion; areithiau gwleidyddol; rhyngweithiadau ffurfiol ac anffurfiol yn yr Eisteddfod Genedlaethol	250,000 yr un	
Ysgrifenedig	Gwaith ysgrifenedig gan ddysgwyr Cymraeg	600,000	4 miliwn
	Llyfrau; papurau bro; dogfennau gwleidyddol; storïau	400,000 yr un	
	Llythyrau a dyddiaduron; traethodau academaidd; gwerslyfrau academaidd; cylchgronau; hysbysebion, taflenni gwybodaeth/cyhoeddusrwydd; llythyrau ffurfiol	290,000 yr un	
	Arwyddion	60,000	
E-iaith	Byrddau trafod; negeseuon e-bost; blogiau	500,000 yr un	2 filiwn
	Gwefannau; trydariadau	300,000	
	Negeseuon testun	200,000	
		10,000,000	

Tabl 2. Fframwaith samplu diwygiedig bras ar gyfer elfen lafar CorCenCC.

Cyd-destunau	% yr is-gorpws	Nifer y geiriau
Cyhoeddus/sefydliadol	10%	400,000
Cyfryngau	15%	600,000
Trafodol	10%	400,000
Proffesiynol	10%	400,000
Addysgegol	10%	400,000
Cymdeithasu	22.5%	900,000
Preifat	22.5%	900,000
	100%	**4,000,000**

Tabl 3. Fframwaith samplu diwygiedig bras ar gyfer elfen ysgrifenedig CorCenCC.

Ffynonellau	% yr is-gorpws	Nifer y geiriau
Llyfrau	41.75%	1,670,000
Cylchgronau, papurau newydd, cyfnodolion	19.25%	770,000
Deunydd amrywiol	39%	1,560,000
	100%	4,000,000

Tabl 4. Fframwaith samplu diwygiedig bras ar gyfer elfen e-iaith CorCenCC.

Ffynonellau	% yr is-gorpws	Nifer y geiriau
Blog	30%	600,000
Gwefan	30%	600,000
Negeseuon e-bost	20%	400,000
Negeseuon testun electronig byr	20%	400,000
	100%	**2,000,000**

Er bod y fframwaith samplu'n gweithredu fel offeryn amcangyfrif ar gyfer casglu data – 'delfryd' fel petai – prin iawn y mae corpws gorffenedig yn efelychu cyfansoddiad y fframwaith samplu (gweler Hawtin, 2018 am drafodaethau pellach ar hyn). Cafodd amrywiaeth o ffactorau ddylanwad ar gyfansoddiad terfynol y corpws, gan gynnwys pa mor hygyrch yr oedd unigolion penodol a/neu fathau o ddata, caniatadau, a materion mwy ymarferol yr oedd yn ymwneud â faint o amser mae'n ei gymryd i brosesu mathau penodol o ddata, a'r graddau y gellir rhagweld hyn. Unwaith roedd y ffactorau hyn i gyd wedi chwarae eu rhan, roedd cyfansoddiad y corpws fel a ganlyn (Tabl 5):

Tabl 5. Cyfansoddiad terfynol CorCenCC (bras).

LLAFAR			
llafar_cyd-destun	**Nifer y testunau**	**Nifer y geiriau**	**Cyfanswm**
darllediad	564	750,078	1,332 testun 2,860,095 geiriau
pedagogaidd	136	296,709	
preifat	92	240,719	
proffesiynol	80	477,983	
cyhoeddus neu sefydliadol	137	433,361	
cymdeithasol	131	456,487	
trafodol	191	204,758	
YSGRIFENEDIG			
ysgrifenedig_genre	**Nifer y testunau**	**Nifer y geiriau**	**Cyfanswm**
cyfnodolyn	10	304,447	707 testun 3,934,082 geiriau
llyfr	137	1,928,582	
traethodau_gwaith cwrs_ac_arholiadau	31	26,047	
taflen_dogfen_cyhoeddiad	338	800,030	
llythyr	53	12,873	
cylchgrawn	80	329,203	
amrywiol	5	8,251	
cylchlythyr	33	78,803	
papur_bro	13	117,334	
traethawd hir	4	328,512	
IAITH ELECTRONIG			
eiaith_genre	**Nifer y testunau**	**Nifer y geiriau**	**Cyfanswm**
blog	48	2,345,909	9,397 testun 4,402,003 geiriau
e-bost	781	141,554	
neges destun	8,487	93,541	
gwefan	81	1,820,999	
	11,432	**11,196,180**	

Mae'r corpws yn cynnwys dros 11,000,000 o eiriau, ond mae'r cyfansoddiad wedi newid fel bod yr elfen lafar yn cynnwys ychydig dros 2,800,000 o eiriau. Er bod yr is-gorpws hwn ychydig yn llai na'r hyn a gynlluniwyd yn wreiddiol, *ynddo'i hun*, dyma'r corpws mwyaf o safbwynt y Gymraeg llafar a siaredir yn naturiol. Am ddadansoddiad manwl o'r corpws (gan gynnwys cyd-destunau, genres a phynciau penodol ynghyd â chategorïau metadata demograffig a'u diffiniadau) gweler Knight ac eraill 2021a (adran 10.2.).

Dylid nodi fod yr offer ymholi corpws ar-lein yn rhoi cyfanswm o 14,338,149 o **docynnau** yn y corpws ac yn gwneud cyfrifiadau ar sail y gwerth hwnnw. Tocynnau yw'r uned leiaf a gynhwysir mewn corpws, sy'n cynnwys geiriau (h.y. eitemau sy'n dechrau gyda llythyren o'r wyddor) a ffugeiriau (h.y. eitemau sy'n dechrau gyda nod nad yw'n llythyren o'r wyddor). Felly, mae corpora bob amser yn cynnwys mwy o docynnau na geiriau. Seilir y gwerthoedd a drafodir yn y bennod hon ar eiriau yn unig gan fod modd dadlau bod hyn yn rhoi cyfrif mwy cywir o'r **unedau ystyr** a gynhwysir yn y corpws.

Cyrchu data

Wrth recriwtio cyfranwyr data llafar, y nod oedd sicrhau y byddai pob ardal yng Nghymru'n cael ei chynrychioli. Cyrchwyd data llafar drwy dau brif ddull: (i) recriwtio'r cyfranogwyr y byddai'n cael eu recordio a (ii) recriwtio'r cyfranogwyr a fyddai'n cyfrannu data llafar drwy ap CorCenCC (gweler 3.3. hefyd). Roedd cwmpas (i) nid yn unig yn golygu cynorthwywyr ymchwil a recordiodd siaradwyr yn y maes, ond hefyd cyfranogwyr yn recordio'u hunain yn ystod rhyngweithiadau amrywiol. Cafodd hyn ei hwyluso drwy rwydwaith o 'ymgyrchwyr' lleol (*animateuriaid* ieithyddol gweithredol ym meysydd a dargedwyd) neu'r Mentrau Iaith (mae Menter Iaith yn gysylltiedig â phob awdurdod lleol yng Nghymru, h.y. sefydliad sy'n seiliedig yn y gymuned sy'n ymrwymedig i godi proffil y Gymraeg drwy fentrau iaith lleol). Cyflawnwyd y broses recriwtio ar gyfer (ii) drwy roi cyhoeddusrwydd i'r ap (er enghraifft, drwy'r cyfryngau cymdeithasol, siarad ar y teledu a deunydd cyhoeddusrwydd) er mwyn gwneud popeth posibl i gyrraedd carfan wahanol o gyfranogwyr a fyddai'n recordio ar eu pen eu hunain ac mewn meysydd mwy preifat. Rhoddodd digwyddiadau mawr yng Nghymru, fel yr Eisteddfod Genedlaethol a Thafwyl, gyfleoedd i'r tîm gyrraedd trawstoriad o gyfranogwyr yn ogystal â chodi ymwybyddiaeth o'r prosiect yn gyffredinol.

Roedd yn her recriwtio pobl i gyfrannu gan ddefnyddio ap ffôn CorCenCC. Roedd yr ap ar gael ar iOS ac Android a thrwy ryngwyneb ar y we (i ddarparu ar gyfer y bobl hynny nad oedd ffôn symudol yn hygyrch iddynt), a chynhyrchodd ymgyrchoedd yn y cyfryngau lawer o frwdfrydedd chychwynnol, e.e. ei drafod ar raglenni teledu fel *Prynhawn Da* ar S4C, ac ar y radio drwy gyfrwng y Gymraeg a'r Saesneg, a thrwy ddigwyddiadau ymgysylltu lleol. Fodd bynnag, ni wnaeth hyn olygu bod llawer o bobl wedi defnyddio'r ap. Roedd adborth gan bartneriaid yn y Mentrau Iaith yn awgrymu y gallai pobl fod yn orbryderus y byddai modd eu hadnabod (er gwaethaf ymdrechion i roi sicrwydd y byddent yn cael eu hanonymeiddio), oherwydd bod y gymuned ieithyddol yn gymharol fach.

Roedd deunydd hyrwyddo (yr oedd wedi'i anelu at annog cyfranogiad, ond a oedd hefyd yn ffordd effeithiol o godi ymwybyddiaeth o'r prosiect) yn cynnwys ysgrifbinnau, matiau diod, taflenni a thaflenni gwybodaeth o faint cerdyn post. Cafodd masgot 'answyddogol' – yn seiliedig ar gath o'r enw Cor-pws – ei ddylunio er mwyn hwyluso cyfranogiad gan bobl o dan 18 oed, a buodd yn boblogaidd gan gyfranwyr o bob oedran.

Sefydlwyd cyfrifon Facebook a Twitter ar gyfer CorCenCC yn ystod ychydig fisoedd cyntaf y prosiect er mwyn gwella'r broses o recriwtio cyfranogwyr a'r cyfraniad ganddynt.

O safbwynt data ysgrifenedig, arweiniodd y cydberthynas da a sefydlwyd â chyhoeddwyr Cymraeg fel gwasg y Lolfa ar ddechrau'r prosiect at ymgorffori llawer o nofelau a llyfrau cyfoes yn y corpws. Mae'r papurau bro lleol yn ffynhonnell unigryw o ddata ysgrifenedig yn y Gymraeg (h.y. papurau newydd Cymraeg ar gyfer y gymuned leol). Penderfynwyd gweithio â'n cysylltiadau Menter Iaith lleol er mwyn casglu'r rhain. O ganlyniad i'n hymgysylltiad â rhanddeiliaid eraill y prosiect yn ystod cam cynllunio'r prosiect, cafodd data ei gasglu'n eithaf cyflym, er enghraifft gwaith samplu o gyfnodolyn academaidd Cymraeg *Gwerddon* drwy'r Coleg Cymraeg Cenedlaethol, ac adnoddau addysgegol / papurau arholiad ail iaith i oedolion drwy Gyd-bwyllgor Addysg Cymru.

O safbwynt data e-iaith, nid oeddem yn gallu casglu data o gyfrifon Twitter na Facebook oherwydd problemau'n ymwneud â chyfyngiadau perchenogaeth, ond cydweithiodd perchenogion gwefannau ac awduron blogiau'n hael, a churwyd y targedau gan gryn tipyn. Bu'n anoddach cipio negeseuon testun electronig byr (mewn modd tebyg i gasglu data drwy'r ap, ac am yr un rhesymau), ond bu'n haws casglu cyfraniadau drwy rif WhatsApp neilltuedig. Yn yr un modd, roedd casglu negeseuon e-bost personol yn her, ond roedd hi'n haws casglu e-ohebiaeth o weithleoedd. Roedd contractau â BBC Cymru/Wales ac S4C wedi golygu y gallai samplau o ddeunydd teledu a radio cyfoes, gan gynnwys podlediadau, gael eu cynnwys. Mae'n werth nodi bod y cydberthnasau gwaith cadarn a sefydlwyd â'r sefydliadau hyn wedi golygu eu bod hefyd wedi cyflenwi negeseuon e-bost o'r gweithle, newyddlenni ac ati ar gyfer y deunydd ysgrifenedig.

Cafwyd cymeradwyaeth foesegol ar gyfer pob agwedd ar gasglu data gan bob un o'r pedair prifysgol a fu ynghlwm wrth y prosiect. Roedd y ffurflenni caniatâd a gafodd eu llofnodi gan gyfranogwyr yn cynnwys eu cydsyniad ar gyfer casglu metadata pwysig (e.e. oedran, rhyw, lleoliad daearyddol) a oedd yn angenrheidiol ar gyfer y corpws. Gweler Knight ac eraill, 2021a, am drafodaeth fanwl ar yr ystyriaethau moesegol / heriau a wynebwyd wrth adeiladu CorCenCC.

Trawsgrifio

Er bod confensiynau trawsgrifio wedi cael eu creu ar gyfer y Gymraeg ar gyfer prosiectau eraill (e.e. Deuchar et al., 2014), penderfynwyd creu set o gonfensiynau bwrpasol ar gyfer trawsgrifio data CorCenCC. Gwnaeth y rhain ein galluogi i adlewyrchu'n llawn y sbectrwm cyfan o amrywiaeth mewn tafodieithoedd/cyweiriau a gipiwyd yn ein data llafar (gan ei wneud yn fwy defnyddiol i ymchwilwyr academaidd) yn ogystal a chynrychioli iaith lafar y cyfranogwyr eu hunain mewn modd mwy cywir. Yn benodol, oni chynrychiolwyd y gwahaniaethau'n gywir, ni fyddai'n bosibl cipio amrywiaethau yn y Gymraeg, na'r pellter a geir rhwng llawer o amrywiadau llafar a'r Gymraeg ysgrifenedig safonol. Rhoddwyd cyfarwyddyd i drawsgrifwyr beidio â chywiro patrymau llafar y gellid eu hystyried yn ansafonol neu a oedd yn cynnwys newid cod (h.y. newid rhwng ieithoedd gwahanol yn ystod achos o gyfathrebu). Roedd Y Tiwtiadur (y pecyn cymorth pedagogaidd – gweler WP4) yn datgan y dylem allu nodi iaith anweddus fel y gellid ei hepgor o gymwysiadau corpws a ddefnyddir â phlant mewn modd systematig, er enghraifft, felly rhoddwyd cyfarwyddyd i drawsgrifwyr nodi achosion posibl ar gyfer eu hadolygu.

Roedd recriwtio trawsgrifwyr yn her barhaus. Fel y trafodwyd uchod, safodd pob darpar drawsgrifydd brawf cychwynnol lle'r oedd angen iddo drawsgrifio darn byr yn unol â chonfensiynau trawsgrifio CorCenCC. Ceir trosolwg manwl o'r penderfyniadau a wnaethpwyd wrth ddatblygu'r confensiynau trawsgrifio ar gyfer CorCenCC, a'r rhesymeg drostynt, yn Knight ac eraill, 2021a.

3.4. WP1: Cyfraniadau allweddol

Prif gyfraniad WP1 yw'r data o 11 miliwn o eiriau sy'n ffurfio craidd y corpws. Yn ogystal, mae'r amrediad canlynol o adnoddau, y cafodd eu creu fel rhan o'r broses o gynhyrchu data, yn helpu i gyflawni un o'r nodau a bennwyd ar gyfer prosiect CorCenCC: cynyddu'r gallu ac ymestyn y rhyngwyneb rhwng y Gymraeg (ac ieithoedd eraill sydd wedi'u lleiafrifo o gwmpas y byd, yn eu tro) a disgyblaeth ieithyddiaeth gymhwysol (gan gynnwys, ieithyddiaeth corpws, sosioieithyddiaeth, a chynllunio a pholisi ieithyddol, yn benodol):

- Fframwaith samplu ar gyfer creu corpws cyffredinol o iaith leiafrifol;
- Diffiniad o 'iaith anweddus' sy'n addas ar gyfer cyd-destun iaith leiafrifol lle mae'r holl siaradwyr yn ddwyieithog;
- Set bwrpasol o gonfensiynau trawsgrifio y gellir eu cymhwyso i Gymraeg lafar cyfoes;
- Tîm o gynorthwywyr ymchwil sy'n siarad Cymraeg, sydd wedi derbyn hyfforddiant ar egwyddorion creu corpws, sydd wedi cael y cyfle i weithio gydag arbenigwyr rhyngwladol ar draws y byd, ac sy'n gallu cymhwyso'u sgiliau i brosiectau yn y dyfodol.

Cafodd y gwaith a wnaed gan dîm WP1 ei rannu yn ystod amrediad o gynadleddau rhyngwladol a chenedlaethol, a cheir mwy o fanylion o safbwynt cyfraniadau damcaniaethol/methodolegol CorCenCC yn y cyhoeddiadau (e.e. Knight ac eraill, 2021a, Knight ac eraill, 2021b).

3.5. WP1: Cymwysiadau ac effaith

Mae'r broses o gynllunio a gweithredu'r meysydd craidd o waith o fewn WP1 yn cynnig templed i'r bobl hynny sy'n ymchwilio i ieithoedd sydd wedi'u lleiafrifo neu ieithoedd lleiafrifol eraill. Cadarnhaodd adborth a gafwyd yn ystod cynadleddau rhyngwladol fod gan ymchwilwyr mewn lleoliadau eraill ymwybyddiaeth o bosibiliadau trosglwyddo methodoleg a phrosesau CorCenCC i brosiectau ieithyddol eraill (e.e. cynllunio corpws ar gyfer yr Wyddeleg a Malteg yn ystod cynadleddau diweddar).

Bydd y corpws a gasglwyd dan nawdd WP1 yn hwyluso ymchwil academaidd i dueddiadau cyfoes yn y Gymraeg. Mae'r mathau o gwestiynau y gellid mynd i'r afael â nhw, o gofio'n protocolau casglu a thrawsgrifio data, yn cynnwys y canlynol:

- A yw patrymau treiglo'n newid yn y Gymraeg cyfoes ac, os ydynt, pa feysydd a genre sydd ar flaen y gad a pha rai sydd o'r tu ôl?
- Sut mae'r ffurfiau a ddefnyddir mewn e-iaith Gymraeg yn cymharu â'r rheiny a geir mewn iaith lafar ac ysgrifenedig (ac amrywiadau gwahanol ohonynt)?

- Beth mae'r corpws yn ei ddweud wrthym am gyflwr presennol y tafodieithoedd daearyddol Cymraeg 'traddodiadol' ac am amrywiadau newydd sy'n dod i'r amlwg, gan gynnwys eu dosbarthiad cymdeithasol?
- Pa eiriau ac ymadroddion, nad oeddent wedi'u cofnodi fel rhan o'r Gymraeg yn ffurfiol yn flaenorol, a geir yn y corpws, ac a ellir nodi unrhyw dueddiadau?
- Pa mor gyffredin yw'r achosion o newid cod mewn data llafar ac e-iaith, a beth sydd fel petai'n ei achosi?

Bydd CorCenCC yn datgelu llawer mwy am fywiogrwydd a bywioldeb presennol y Gymraeg. Bydd yn rhaid i ni fod yn barod i dderbyn y caiff elfennau o'r datgeliadau hyn eu croesawu ac y caiff elfennau eraill eu gwrthod gan y gymuned ehangach o siaradwyr Cymraeg. Yn y pen draw, pe byddai'n ddymuniad gan y gymuned, gall CorCenCC fod yn gyfrwng ar gyfer trafodaeth sy'n seiliedig ar dystiolaeth ar yr hyn sy'n 'Gymraeg safonol' – neu a allai fod – yn ystod yr unfed ganrif ar hugain.

4. Pecyn Gwaith 2: Datblygu set dagiau ar gyfer rhannau ymadrodd a'i defnyddio i dagio data WP1

4.1. WP2: Disgrifiad

Diben WP2 oedd gweithredu modd priodol o nodi a labelu ('pennu tagiau') rhannau ymadrodd (e.e. enw, berf, ac is-deipiau o gategorïau o'r fath) a oedd yn nodweddiadol o'r iaith a gasglwyd yn WP1, fel y gellid chwilio a dadansoddi'r corpws yn y ffyrdd amrywiol y byddai'n ofynnol gan ddefnyddwyr yn y dyfodol. Yr offerynnau gofynnol oedd tagiwr rhannau ymadrodd a set dagiau. Cafodd adnodd a oedd eisoes yn bodoli, sef Autoglosser Bangor (Donnelly a Deuchar, 2011), ei ddefnyddio fel y man cychwyn, oherwydd ei ddibynadwyedd a'i addasrwydd profedig ar gyfer pennu tagiau i'r Gymraeg. Dan arweiniad arweinwyr WP2, sef Knight a'r ymgynghorydd prosiect Donnelly (ieithydd cyfrifiadurol sydd wedi gweithio'n agos ar ddatblygu corpysau Cymraeg), cymhwysodd Neale (un o gynorthwywyr ymchwil y prosiect) Autoglosser Bangor i'r set o ddata WP1 a oedd yn dod i'r amlwg, a nododd lle'r oedd angen addasu a mireinio'r offerynnau – mewn perthynas, yn bennaf, â phennu tagiau i ffynonellau iaith lafar ac e-iaith, a'r amrediad ehangach o ran amrywiad rhanbarthol a genre. Cafodd yr addasiadau eu cymhwyso yn ystod camau diweddaraf y prosiect gan Tovey-Walsh (myfyrwraig PhD y prosiect), gan arwain at dagiwr a set dagiau CorCenCC ei hun, sef CyTag (gweler 4.3. isod).

4.2. WP2: Amcanion

Roedd nodau ac amcanion WP2 fel a ganlyn:

- adeiladu tagiwr rhannau ymadrodd y Gymraeg a'i hyfforddi
- datblygu set dagiau briodol
- pennu tagiau i'r holl ddata

4.3. WP2: Cyflawniadau

Cafodd meddalwedd bwrpasol CorCenCC ar gyfer pennu tagiau i rannau ymadrodd, CyTag, ei datblygu yn ystod deunaw mis cyntaf y prosiect, a chafodd ei rhyddhau yn gyhoeddus ym mis Mawrth 2018. Mae gwerthusiadau a chymwysiadau CyTag hyd yn hyn yn nodi ei bod yn gyflawn ac yn gadarn a'i fod yn gweithio'n dda. Mae CyTag yn liferu deunyddiau ffynhonnell agored i helpu gyda'r broses o benderfynu ar rannau ymadrodd. Mae'n gweithio yn bennaf drwy ddefnyddio'r wybodaeth sydd yng ngeiriadur Donnelly, *Eurfa* – y geiriadur Cymraeg ffynhonnell agored mwyaf sydd ar gael yn rhad ac am ddim (Donnelly, 2013a) – i lunio rhestr o dagiau posibl ar gyfer pob gair mewn testun Cymraeg. Ategir hyn gan restrau penodol o enwau lleoedd, enwau cyntaf a chyfenwau sydd wedi'u dethol o ddata Wikipedia.

Unwaith y mae rhestr o eiriau wedi'i llunio, gellir cymhwyso set o reolau pwrpasol i fireinio'r rhestr o dagiau posibl a bennwyd ar gyfer pob gair – yn seiliedig ar y tagiau neu'r nodweddion a bennwyd i'r geiriau o'i gwmpas – nes cyrraedd yr un cywir. Er enghraifft, gall y ffurf 'yn' olygu 'in', ond mae ganddi hefyd ddwy rôl fel geiryn gramadegol. Mewn un rôl, mae'n trosi ansoddair yn adferf (e.e. *'yn dda'*, o 'da'). Yn y llall, mae'n cael ei gysylltu â'r ferf 'bod' (un ffurf ohoni yw 'mae') er mwyn cyflwyno berf, enw neu ddibeniad ansoddeiriol (e.e. 'mae'r llyfr yn dda'). Fel y gwelir o'r enghreifftiau hyn, gall 'yn dda' olygu 'good' a 'well', ond yn y ddwy achos byddai 'yn' yn cael ei ddosbarthu'r un ffordd. Roedd angen i'r tagiwr allu gwahaniaethu rhwng 'yn' fel arddodiad ac 'yn' fel geiryn, ac mae'n gwneud hynny yn y ffordd ganlynol: Yn y frawddeg 'mae Cymru yn wlad Geltaidd', mae CyTag yn pennu tag cywir i 'yn' fel geiryn sy'n cyflwyno dibeniad, oherwydd bod 'mae' o'i flaen, ac oherwydd bod 'wlad' yn dreiglad meddal o 'gwlad'; mae gennym reol i ddewis tag y geiryn dibeniadol ar gyfer 'yn' lle bo'r gair canlynol yn enw sydd wedi'i dreiglo'n feddal. (Pan fo 'yn' yn golygu 'in', caiff hyn ei ddilyn gan dreiglad trwynol.) Dangosodd ganlyniadau'r broses werthuso fod CyTag yn cyflawni lefel o gywirdeb o dros 95%, sy'n gymharol â'r enghreifftiau gorau o dagwyr rhannau ymadrodd ar gyfer ieithoedd eraill. Ar hyn o bryd, mae CyTag yn cynnwys rhannwr testun, holltwr brawddegau, tocynnwr, a'r tagiwr rhannau ymadrodd ei hun. Ceir gwefan CyTag yn: http://cytag.corcencc.org.

Mae set dagiau rhannau ymadrodd CyTag yn cynnwys 145 o dagiau manwl, h.y. tagiau 'cyfoethog', sydd wedi'u mapio'n 13 categori sy'n cydymffurfio â Safonau'r Grŵp Cynghori Arbenigol ar Beirianneg Ieithyddol (EAGLES, 1996). Mae'r tagiau hyn yn cynnwys categorïau cystrawennol pwysig (e.e. enw, bannod, arddodiad, berf ac ati) yn ogystal â dau gategori sy'n cynrychioli geirynnau sy'n 'unigryw' i'r Gymraeg a ffurfiau 'eraill' fel byrfoddau, acronymau, symbolau, digidau ac ati. Mae'r set lawn o 145 o dagiau yn cwmpasu morffoleg y Gymraeg yn seiliedig ar genedl (gwrywaidd neu fenywaidd), rhif (unigol neu luosog), person (cyntaf, trydydd ac ati) a'r amser (gorffennol, presennol, dyfodol ac ati). Mae'r tagiau eu hunain wedi'u hamgodio'n Gymraeg.

Ceir y set dagiau lawn yn: https://cytag.corcencc.org/tagset?lang=cy. Gweler Neale et al., 2018 am drosolwg technegol trylwyr o CyTag a gwerthusiad manwl o'i chywirdeb.

Un o fanteision pwysig dull CorCenCC o safbwynt datblygu meddalwedd sy'n pennu tagiau yw'r modd y mae'n defnyddio'r nifer lleiaf o reolau sy'n hawdd eu cymhwyso i wybodaeth ac adnoddau sydd eisoes yn bodoli. Mae modd trosglwyddo'r dull hwn i ieithoedd y mae data hyfforddi sydd wedi'i anodi ymlaen llaw amdanynt yn brin, gan ei wneud yn werthfawr ar gyfer cipio nodweddion ieithoedd lleiafrifol.

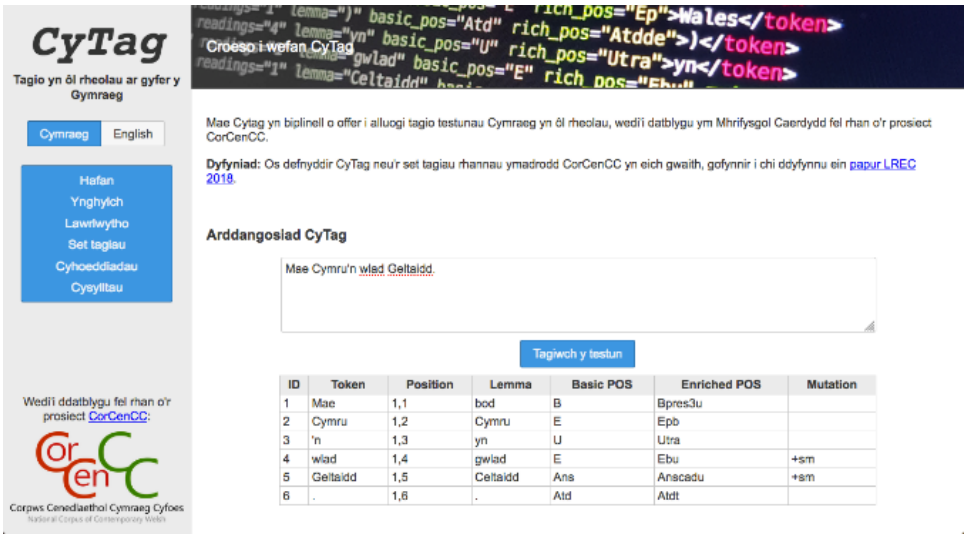
4.4. WP2: Cyfraniadau allweddol

Prif gyfraniad gwaith WP2 oedd creu offerynnau meddalwedd ac adnoddau ieithyddol sydd ar gael yn rhad ac am ddim ac sy'n ymestyn yr adnodd sydd eisoes yn bodoli ar gyfer dadansoddi, a chloddio testun, y Gymraeg. Yn benodol, gwnaethom greu'r canlynol:

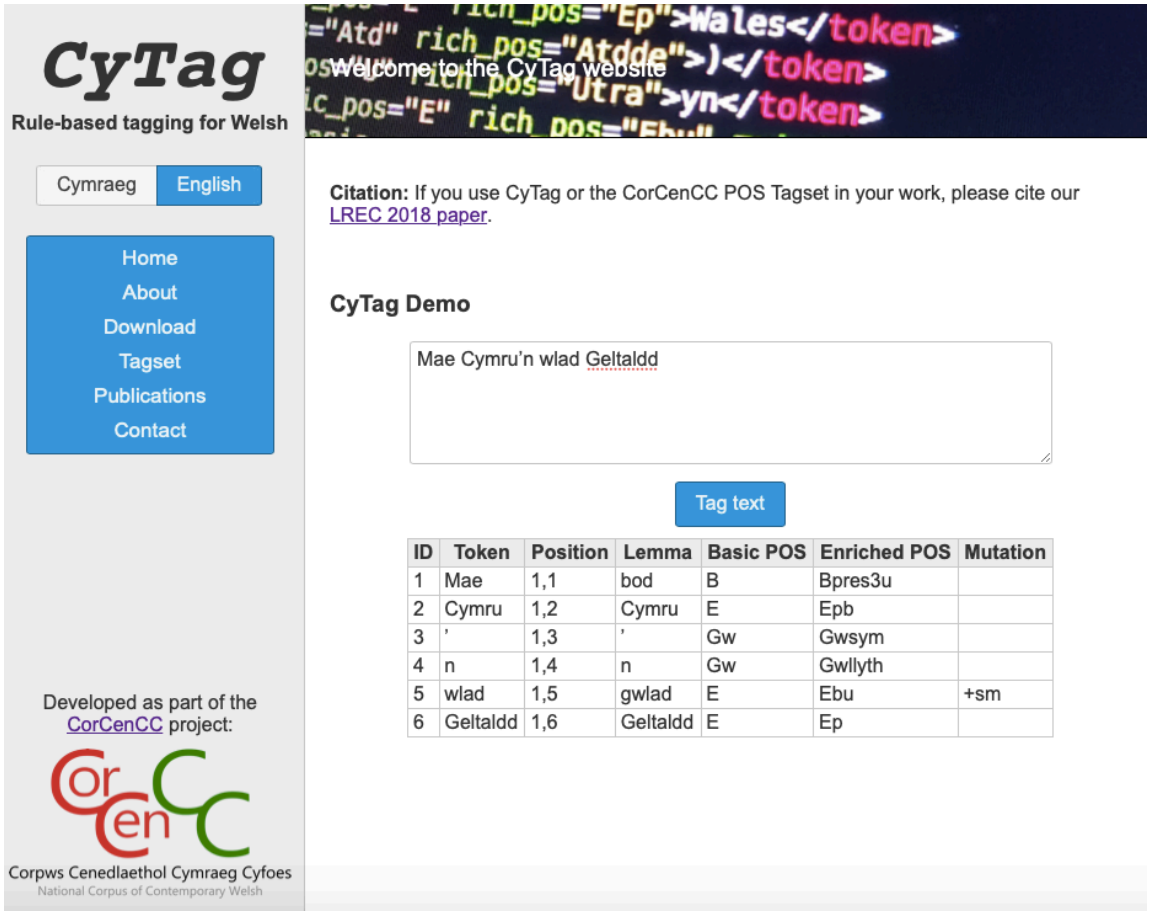
- Set dagiau tagiwr rhannau ymadrodd CorCenCC (https://cytag.corcencc.org/tagset?lang=cy), gyda:
    - 145 o dagiau rhannau ymadrodd 'cyfoethog'
    - 13 categori 'sylfaenol' sy'n cydymffurfio ag EAGLES

  Gellir mabwysiadau'r set dagiau hon fel rhestr safonol (h.y. set o gonfensiynau) a/neu ychwanegu ati a/neu ei haddasu ymhellach gan ddefnyddwyr yn y dyfodol wrth bennu tagiau i setiau data'r Gymraeg.
- Corpws gwerthuso o 'safon aur'. Mae corpws safon aur yn un sydd wedi'i anodi â llaw a'i wirio gan unigolion lluosog. Mae hyn, mewn effaith, yn darparu model sy'n gallu hyfforddi'r dull awtomatig (cyfrifiadurol) a'i werthuso. Mae'r corpws safon aur hwn hefyd wedi'i ryddhau er mwyn i ymchwilwyr eraill allu ei ddefnyddio yn y gwaith o ddatblygu eu hofferynnau eu hunain. Mae'n cynnwys y canlynol:
    - 611 o frawddegau
    - 14,876 o docynnau
- Gwefan CyTag: (https://cytag.corcencc.org)
  Mae'r wefan hon yn cynnwys gwybodaeth ddwyieithog (Cymraeg a Saesneg) am y tagiwr, gan gynnwys arddangosiad gweithiol (gweler Ffigur 1a ac 1b am sgrinluniau yn Gymraeg a Saesneg yn ôl eu trefn). Gall defnyddwyr ddefnyddio'r tagiwr drwy'r wefan hon a'i ddefnyddio i bennu tagiau i'w data eu hunain.
- CyTag ar Github (https://github.com/CorCenCC/CyTag)
  Mae'r tagiwr ffynhonnell agored hefyd ar gael i bawb (h.y. ar gael fel meddalwedd rhad ac am ddim, dan delerau fersiwn 3 Trwydded Gyhoeddus Gyffredinol GNU) drwy wefan GitHub. Yma, gall defnyddwyr eto lawrlwytho'r tagiwr a'i ddefnyddio. Gallant hefyd wella'r tagiwr a rhannu fersiynau a ddiweddarwyd â defnyddwyr eraill y dyfodol.

Ffigur 1a. Sgrinlun o ryngwyneb arddangosol ar-lein CyTag yn Gymraeg



Ffigur 1b. Sgrinlun o ryngwyneb arddangosol ar-lein CyTag yn Saesneg (ar gael yn https://cytag.corcencc.org)



Mae fersiwn gyfredol CyTag yn cynnwys y gwelliannau canlynol a wnaed gan Tovey-Walsh:

- fersiwn ddiwygiedig o'r tocyneiddiwr
- diweddariad i'r modd yr ymdrinnir â'r treigladau sy'n gwella sut mae'n medru dal treiglad mewn geiriau byrrach ac enwau priod
- cynnwys rhai geiriau gyda phrif gymeriad â llythyren fach yn y lecsicon (e.e. 'cymraeg')
- cynnwys geiriau Saesneg amlder uchel fel bod modd i'r tagiwr eu hadnabod (fel geiriau nad ydynt yn rhai Cymraeg) sy'n lleihau nifer y geiriau sy'n cael eu tagio'n 'anhysbys'
- cynnwys banodolion a rhagenwau amhendant yn y set o dagiau (e.e. 'neb', 'pob') a'r modiwl tagio gan ychwanegu'r rheolau hyn i ramadeg cyfyngu CyTag

4.5. WP2: Cymwysiadau ac effaith

Mae CyTag ar gael fel tagiwr ynddo'i hun. Mae hyn yn golygu y gall unrhyw un sy'n gwybod sut i ddefnyddio tagiwr ei ddefnyddio i bennu tagiau i'w ddata ei hun. Mae CyTag wedi'i chymhwyso i CorCenCC yn ei gyfanrwydd, fel nad oes angen i ddefnyddwyr wybod sut i ddefnyddio tagiwr er mwyn cael gwybodaeth bwysig am yr iaith. Gallai defnyddwyr o'r fath gynnwys ymchwilwyr, athrawon iaith, datblygwyr technoleg a geiriadurwyr. Gall defnyddwyr eraill hefyd wella CyTag yn y dyfodol.

5. Pecyn Gwaith 3: Cwmpasu a datblygu tagiwr semantig ar gyfer y Gymraeg a'i ddefnyddio i bennu tagiau semantig i ddata WP1

5.1. WP3: Disgrifiad

Cafodd Pecyn Gwaith 3 (WP3) ei arwain gan Rayson, gyda Piao yn gweithredu fel cysylltai ymchwil hyd at fis Awst 2018 (pan ymgymerodd â swydd darlithio amser llawn), pan ymunodd Ezeani â'r tîm. Defnyddiodd tîm WP3 yr arbenigedd yn y Gymraeg a oedd ar gael ar draws gymuned gyfan prosiect CorCenCC yn ôl yr angen. Prif elfen gwaith ymchwil WP3 oedd cynllunio a datblygu'r system feddalwedd ar gyfer anodiadau semantig y Gymraeg a fyddai'n llywio'r adnoddau ieithyddol cysylltiedig.

5.2. WP3: Amcanion

Roedd gan WP3 bedwar nod ac amcanion fel a ganlyn:

- cynllunio set dagiau semantig newydd ar gyfer y Gymraeg
- datblygu'r system anodi awtomatig
- rhoi prawf ar ddulliau torfol ar gyfer pennu tagiau semantig
- pennu tagiau semantig i'r holl ddata

Roedd y gwaith ymchwil a gyflawnwyd yn WP3 yn adeiladu ar gwerth 30 blynedd o waith dadansoddi semantig awtomatig ym maes ieithyddiaeth corpws ac ieithyddiaeth gyfrifiadurol yng nghanolfan ymchwil Prifysgol Caerhirfryn. Roedd y gwaith helaeth a oedd eisoes wedi'i wneud ar ieithoedd eraill (Ffinneg, Rwsieg, Tsieineeg, Iseldireg, Eidaleg, Portiwgaleg, Sbaeneg, Maleieg) o werth arbennig ar gyfer CorCenCC.

Fel rhan o brosiect CorCenCC, roedd angen i ni ailwerthuso set dagiau System Dadansoddi Semantig UCREL (USAS) er mwyn darparu ar gyfer nodweddion arbennig y Gymraeg, a gofynion ymarferol y gwaith o ddatblygu'r pecyn cymorth pedagogaidd (WP4) a defnyddwyr terfynol y corpws (WP5). Gwnaethom hefyd anelu at ddatblygu algorithmau a dulliau newydd i ddynodi meysydd semantig a oedd yn briodol i'r cyd-destun i unedau geiriadurol y Gymraeg, boed y rheiny'n eiriau unigol neu'n ymadroddion â geiriau lluosog. Hefyd, i ategu at y dull torfol o gasglu data'r corpws yn WP1, gwnaethom anelu at beilota dulliau cyfrannu torfol ar gyfer dynodi meysydd semantig er mwyn ymestyn y geiriaduron semantig gwaelodol a'u galluogi i adlewyrchu'r hyn a ddehonglwyd gan siaradwyr Cymraeg.

5.3. WP3: Cyflawniadau

Y dasg gyntaf oedd creu set dagiau USAS ar gyfer y Gymraeg. Cyflawnwyd hyn drwy ddefnyddio dulliau a oedd wedi'u datblygu yn ystod prosiect blaenorol a ariannwyd gan y Cyngor Ymchwil i'r Celfyddydau a'r Dyniaethau (AHRC), sef SAMUELS (AH/L010062/1), a diweddaru'r fframwaith Java ar gyfer y Gymraeg. Roedd y gwaith dadansoddi semantig newydd yn y Gymraeg yn cynnwys y camau canlynol.

Ymgymerwyd â'r gwaith o fapio meysydd semantig draw oddi wrth y fframwaith amlieithog a oedd eisoes yn bodoli, fel y gellid adolygu ei addasrwydd ar gyfer y Gymraeg. Roedd yr ystyr posibl a ddynodwyd i dagiau wedi deillio'n awtomatig, yn y lle cyntaf felly, o'r broses o drosi geiriaduron Saesneg trwy eiriaduron dwyieithog a chorpysau bach a oedd yn cydredeg, ac yna cawsant eu gwirio gan siaradwyr Cymraeg tîm CorCenCC, a'u haddasu

yn ôl yr angen. Cafodd y set dagiau semantig newydd ar gyfer y Gymraeg ei rhyddhau ym mis Ebrill 2016.

Roedd ein gwaith ymchwil ar gyfrannu torfol yn gofyn am greu dull ac arbrofion a olygodd gael cyfranogiad gan ddefnyddwyr Amazon Mechanical Turk (AMT) er mwyn gwneud ymchwil i'r graddau y gallai siaradwyr Cymraeg amhroffesiynol nad oeddent wedi'u hyfforddi greu cofnod geiriadurol semantig o safon eithaf uchel drwy ddynodi un neu fwy o feysydd semantig addas, yr oedd wedi'u pennu ymlaen llaw, i air neu ymadrodd o'r corpws. Cafodd y gwaith ymchwil hwn ei gyhoeddi yn Rayson a Piao, 2017 (gweler adran 10.2.).

Roedd angen tagiwr rhannau ymadrodd arnom er mwyn i ni allu gweithio ar WP3, ac er mwyn gwneud cynnydd â hynny cyn bod WP2 wedi'i gwblhau, gwnaethom ddefnyddio tagiwr rhannau ymadrodd dros dro a oedd eisoes yn bodoli, a oedd yn gydran o Becyn Cymorth Iaith Naturiol y Gymraeg (WNLT), fel rhan o'r tagiwr semantig newydd ar gyfer y Gymraeg a gafodd ei greu yn Java (CySemTagger). Yn ddiweddarach, pan oedd allbwn WP2 ar gael, gwnaethom fabwysiadu CyTag. I ddechrau, roedd gwefan rhyngwyneb rhaglennu cymwysiadau (API) (SOAP) ar gael ar gyfrif GitHub UCREL, a alluogodd defnyddwyr i gyrchu'r cymhwysiad. Yn ddiweddarach, roedd rhyngwyneb rhaglennu cymwysiadau newydd (REST) ar gael yn http://ucrel-api.lancaster.ac.uk/, o oedd unwaith eto â'r nod o gynyddu hygyrchedd i'r cymhwysiad hwn.

Cafodd fframwaith Java ei greu gan Piao ar gyfer pennu tagiau i eiriau unigol ac ymadroddion amleiriol, a thrwy ddefnyddio dulliau amrywiol cafodd yr adnoddau ieithyddol eu creu. Mae'r adnoddau terfynol yn cynnwys 143,287 o gofnodion o eiriau unigol a chasgliad o sampl o gofnodion amleiriol, yn ogystal â 329,800 o ffurfdroadau o gorpysau amrywiol. Cafodd fersiwn ar y we o ryngwyneb defnyddiwr graffigol (GUI) y tagiwr semantig Cymraeg ei chreu (gan ddefnyddio rhyngwyneb rhaglennu cymwysiadau SOAP), yn ogystal â fersiwn bwrdd gwaith ohono. Trwy gydweithrediad â'r tîm CorCenCC ehangach, cafodd corpws o safon aur ei wirio â llaw at ddibenion gwerthuso a gwella'r tagiwr (ceir manylion o'r corpws safon aur yn 4.4.).

Er mwyn cwblhau ein gwaith ymchwil ar gyfer y pecyn gwaith hwn, ymgymerodd Ezeani (cynorthwyydd ymchwil) ag arbrawf dysgu aml-dasg er mwyn ymchwilio i ba raddau y gellid defnyddio modelau ymgorffori geiriau o'r radd flaenaf sy'n seiliedig ar fectorau ar gyfer ieithoedd sydd ag adnoddau prin (Cymraeg yn ein hachos ni) gyda modelau rhwydwaith niwral ar gyfer pennu tagiau i rannau ymadrodd a phennu tagiau semantig. Dangosodd ein canlyniadau fod dull o bennu tagiau o'r fath yn cymharu'n dda iawn â'r tagwyr a oedd eisoes yn bodoli (gweler Rayson a Piao, 2017 – adran 10.2). Rydym wedi sicrhau bod ein tagwyr a'n hadnoddau ieithyddol i gyd ar gael ar ffurf ffynhonnell agored gyda thrwyddedau caniatadol.

5.4. WP3: Cyfraniadau allweddol

Un o brif gyfraniadau gwaith ymchwil WP3 yw'r offerynnau meddalwedd a'r adnoddau ieithyddol sydd ar gael yn rhad ac am ddim ac sy'n ymestyn y banc adnoddau ar gyfer dadansoddi, a chloddio testun, y Gymraeg. Mae'r cod Java ar gyfer CySemTagger wedi'i ryddhau ar GitHub ac mae wedi'i ymgorffori yn system anodi a dadansoddi corpws Wmatrix (Rayson ac eraill, 2004). Gwneir defnydd helaeth o'r system hon ym maes ieithyddiaeth corpws, ac mae'n golygu y gall ymchwilwyr ei defnyddio ar gyfer corpysau'r Gymraeg. Yn

gyffredinol, mae gwaith WP3 wedi ymestyn cwmpas y gwaith ymchwil ym meysydd ieithyddiaeth corpysau ac ieithyddiaeth gyfrifiadurol mewn o leiaf dwy ffordd. Yn gyntaf, mae wedi arddangos dull ar gyfer ymestyn technegau dadansoddi semantig i heriau penodol y Gymraeg mewn modd effeithiol. Yn ail, mae wedi dangos y gellir defnyddio dulliau cyfrannu torfol i gyfrannu at y gwaith o ddatblygu adnoddau o'r fath.

5.5. WP3: Cymwysiadau ac effaith

Mae'r tagiwr semantig a ddatblygwyd yn WP3 wedi'i gymhwyso i CorCenCC yn ei gyfanrwydd, gan gynhyrchu cyfoeth o ddehongliadau o'r data a fydd o werth i'r bobl hynny sydd â diddordeb yn y ffordd y defnyddir y Gymraeg cyfoes i gyfleu ystyr ar draws genres, arddulliau a chyfryngau, a sut gellir awtomeiddio prosesau ieithyddol ar gyfer technolegau newydd. Yn ogystal, mae'r egwyddorion gwaelodol a sefydlwyd wrth greu'r CySemTagger yn darparu sylfaen ar gyfer estyniadau pellach yn y dyfodol fel y gellir dadansoddi ieithoedd eraill sydd ag adnoddau prin. Mewn modd tebyg, bydd ymchwilwyr yn gallu ymestyn yr arbrofion dysgu aml-dasg i ieithoedd eraill er mwyn ymchwilio i ba raddau y gall yr ieithoedd hynny fanteisio hefyd, er gwaethaf eu fframweithiau gramadegol gwahanol. Mae'r ffaith fod y CySemTagger wedi'i ymgorffori yn Wmatrix (Piao ac eraill, 2018) yn golygu y gall ymestyn ar draws y gymuned ymchwil ryngwladol, gan alluogi gwaith dadansoddi cynnwys awtomatig ar gorpysau'r Gymraeg y gellid eu casglu gan bobl eraill yn y dyfodol.

6. Pecyn Gwaith 4: Cwmpasu, dylunio ac adeiladu pecyn cymorth pedagogaidd ar-lein

6.1. WP4: Disgrifiad

Cafodd waith WP4 ei arwain gan Thomas a Fitzpatrick, a weithiodd â'r cynorthwywyr ymchwil Needs a J. Davies, ac arweinydd prosiect Knight, a chafwyd cyngor ymgynghorol gan Stonelake (cyd-ymchwilydd), Anthony (ymgynghorydd prosiect – dyluniwr a datblygwr AntConc), Cobb (ymgynghorydd prosiect – datblygwr Compleat Lexical Tutor) ac E. Davies (CBAC).

Mae dysgu/addysgu iaith yn un maes lle gall corpysau fod yn arbennig o addysgiadol. Wrth i athrawon a dysgwyr ddod yn fwy deallus wrth ddefnyddio technoleg, ac wrth i gorpysau barhau i ddatblygu o ran eu maint a'r modd o'u cymhwyso a'u hymarferoldeb, mae dysgu sydd wedi'i lywio gan gorpysau'n ennill ei blwyf yn gyflym o safbwynt lleoliadau dosbarth ac astudiaethau personol. Gellir defnyddio corpysau i amlygu'r geiriau, ymadroddion a phatrymau mwyaf cyffredin mewn iaith benodol. Gallant ddangos pa eiriau sy'n dueddol o fynd gyda'i gilydd, a pha rai sy'n digwydd ym mha fathau o destun (e.e. testunau ysgrifenedig ffurfiol, sgyrsiau llafar, negeseuon e-bost proffesiynol, neu negeseuon testun personol). Gall defnyddwyr corpysau chwilio am eiriau penodol a'u gweld mewn brawddegau enghreifftiol. Gall corpysau, felly, ddarparu ffynhonnell sydd â chyfoeth o iaith i'r dysgwr sy'n arddangos sut mae ei iaith darged yn cael ei defnyddio mewn gwirionedd, ac o fewn meysydd amrywiol.

Un o agweddau arloesol prosiect CorCenCC, a arweiniwyd gan dîm WP4, yw datblygiad cyfres o offerynnau pwrpasol – Y Tiwtiadur – a fwriedir iddo gael ei ddefnyddio

yn ystod dosbarthiadau Cymraeg a'r tu allan iddynt, o ysgolion cynradd i addysg i oedolion. Gyda'i gilydd, mae'r offerynnau hyn yn helpu i arddangos sut y caiff y Gymraeg ei defnyddio mewn gwirionedd mewn pedwar ymarfer arwahanol sy'n seiliedig ar y corpws sy'n defnyddio data a gasglwyd yn WP1 a'r tagwyr / setiau tagiau a ddatblygwyd yn WP2 a WP3. Mae'r offerynnau'n cynnwys y canlynol:

- offeryn Cau Bylchau (Cloze) sy'n galluogi athrawon (neu ddysgwyr, mewn cyd-destunau astudiaethau personol) i ddileu geiriau o unrhyw ddarn o destun yn y corpws, fesul bwlch penodol, er mwyn annog neu asesu medrau deall a strategaethau rhagfynegi
  - Mae'r offeryn hwn yn galluogi defnyddwyr i greu tasg llenwi bylchau gan ddefnyddio darnau o destun yn CorCenCC. Mae'r opsiwn "Math o destun" yn ei gwneud yn bosibl i ddefnyddwyr ddethol genres o destun arbennig (e.e. genres 'blog' neu 'ffuglen llyfrau'). Mae'r gosodiad "Amlder bylchau" yn galluogi defnyddwyr i osod pa mor aml y maent am i fwlch ymddangos, gan ddibynnu ar ba mor anodd y mae angen i'r dasg fod (y gosodiad a argymhellir yw pob seithfed i nawfed gair). Gan ddefnyddio'r opsiwn "Hyd y testun", gall defnyddwyr ddewis gweld sampl ar hap o ddarnau o destun hyd at 100, 200, 300, 400, neu 500 o eiriau o hyd. Wrth glicio "Dechrau", mae panel newydd yn dangos y dasg llenwi bylchau ac mae'r geiriau sydd wedi'u tynnu allan o'r darn o destun yn ymddangos mewn panel ar wahân. Er mwyn cwblhau'r dasg, mae'n ofynnol i ddefnyddwyr ddewis geiriau o'r rhestr a'u teipio i mewn i'r bylchau priodol yn y darn o destun. Pan gaiff "Gwirio" ei glicio, caiff y geiriau sydd wedi'u gosod yn gywir eu hamlygu'n wyrdd, a'r geiriau na osodwyd yn gywir yn goch. Mae argaeledd y darnau o destun o'r corpws yn galluogi athrawon a dysgwyr i gwblhau'r gweithgaredd hwn lawer o weithiau, a cheir cyfleoedd dysgu newydd bob tro.
- offeryn Proffilydd Geiriau sy'n ei gwneud yn bosibl graddio darnau o destun yn ôl pa mor aml mae geiriau'n ymddangos
  - Mae'r offeryn hwn yn pennu proffil i ddarn o destun sydd wedi'i ddethol neu'i greu gan y defnyddiwr yn ôl pa mor aml mae geiriau'n ymddangos. Mae'n ofynnol i ddefnyddwyr gopïo a gludo darn o destun i'r maes "Mewnbynnu testun" neu i deipio testun i'r maes yn uniongyrchol. Wrth glicio ar "Dechrau" mae proffil yn cael ei greu, lle caiff pob gair ei ddynodi i gategori yn ôl pa mor aml y mae'n ymddangos. Mewn panel ar wahân, ceir esboniad o'r canlyniadau. Mae'r colofnau "Lefel" / "Band amlder" yn ymwneud â'r nifer o weithiau y mae gair yn ymddangos yn y 11 miliwn o eiriau a geir yn CorCenCC. Mae geiriau sy'n perthyn i'r band "K1" (y 1,000 uchaf) ymhlith 1,000 o eiriau mwyaf cyffredin y Gymraeg, yn ôl CorCenCC. Fel arfer, y mwyaf o eiriau sydd wedi'u cynnwys mewn darn o destun sy'n perthyn i'r bandiau amlder isaf (e.e. y rheiny o fewn 3,001 – 4,000 (K4), 4,001 – 5,000 (K5) a >5,001 (K6)), y mwyaf heriol y bydd i'r dysgwr ei ddeall. Gellir hefyd pennu proffil i ddarnau o destun a gynhyrchir gan ddysgwyr; mae dysgwyr fel arfer yn casglu mwy o eiriau sydd ag amlder uchel i ddechrau, ac yn meistroli bandiau amlder isel wrth i'w hyfedredd ddatblygu (gweler, er enghraifft, Nation, 2001). Yn y gosodiad diofyn, bydd yr offeryn yn

amlygu geiriau sy'n perthyn i lefelau K1 i K6+. Gall defnyddwyr newid yr offeryn i amlygu geiriau nad ydynt yn perthyn i'r lefelau hyn drwy glicio ar yr opsiwn "Amlygu geiriau nad ydynt yn perthyn i lefel". Gall geiriau sy'n perthyn i'r band 5,001+ gynnwys geiriau sydd wedi'u camsillafu a geiriau o ieithoedd eraill, yn ogystal â geirfa a ddefnyddir yn anaml yn y corpws, neu nad yw wedi'i chipio gan y corpws.

- offeryn Nodi Geiriau sy'n profi gallu dysgwyr i ddyfalu gair o fewn cyd-destun
  - Mae'r offeryn hwn yn arddangos darnau corpws lluosog (llinellau cydgordio) sydd i gyd yn cynnwys gair penodol. Mae'r gair wedi'i guddio, a'r dasg yw nodi'r gair sy'n cyfateb â'r bylchau i gyd. Ceir opsiynau ar gyfer dethol "Band amlder" y gair (K1, K2 a K3), y "Math o air" penodol (e.e. enw, berf) a'r nifer fwyaf o frawddegau i'w harddangos. Er mwyn cynhyrchu'r darnau hyn, mae defnyddwyr yn clicio ar "Dechrau". Yn yr offeryn hwn, yr ymatebion 'cywir' yw'r rheiny sydd wedi'u cynnwys yn CorCenCC; mewn rhai achosion, gallai ymateb gwahanol hefyd fod yn gredadwy o fewn yr iaith ar y cyfan.
- offeryn Creu Tasgau Geiriau sy'n hwyluso gwaith dwys ar eitem benodol o eirfa
  - Mae'r offeryn hwn yn cynhyrchu darnau corpws lluosog (llinellau cydgordio) o CorCenCC sydd i gyd yn cynnwys gair targed a bennir gan y defnyddiwr. Mae defnyddwyr yn teipio'u gair targed i mewn i'r blwch mewnbynnu "Gair". Yna, maent yn dethol sawl darn maent am eu cynhyrchu (uchafswm o 20) gan ddefnyddio'r opsiwn "Uchafswm llinellau". Os yw defnyddwyr am bennu rhan ymadrodd y gair targed (e.e. enw, berf), gallant wneud hynny drwy'r opsiwn "Rhan ymadrodd". Mae'r dasg yn cael ei chreu drwy glicio ar "Dechrau". Mae'r offeryn hwn yn gwneud dau fath o weithgaredd dysgu'n bosibl. Un ohonynt yw arsylwi ar y geiriau sydd o gwmpas y gair penodol. Mae hyn yn helpu'r broses o gasglu strwythurau gramadegol a phatrymau cydleoliad. Y llall yw ffurf fwy firieniol ar yr offeryn Nodi Geiriau uchod, lle gall athro bennu'r gair penodol i'w ddyfalu, yn hytrach na gair a gynhyrchir gan y feddalwedd o set yn unig. Er mwyn hwyluso ail ddefnydd yr offeryn hwn, mae'r gair wedi'i guddio yn y tabl canlyniadau. Mae clicio ar "Dangos" yn datgelu'r gair targed.

Mae'r offerynnau hyn yn ei gwneud yn bosibl i ddefnyddwyr weithio gyda'r iaith mewn sawl ffordd. Er enghraifft, gallant archwilio patrymau cydgordio ar draws sawl cystrawen (e.e. berf + arddodiad, ansoddair + arddodiad, cysylltair (os, fel) + amser dilynol), a dyfalu'r gair coll drwy archwilio a dadansoddi geiriau eraill a ddefnyddir ar y cyd ac yn y cyd-destun yn yr amgylchedd uniongyrchol. Gallant nodi bylchau yn eu geirfa, a rhoi blaenoriaeth i ba eiriau i'w dysgu nesaf. Ochr yn ochr â'r offerynnau ymholi cyffredinol y gellir eu defnyddio yn CorCenCC, mae Y Tiwtiadur yn enghraifft unigryw o ddysgu a yrrir gan ddata (Johns, 1991) ar ffurf addysgeg anwythol a ddefnyddir yn uniongyrchol ac sy'n seiliedig ar gorpws sy'n gallu helpu i ychwanegu at ddysgu'r Gymraeg ar hyd oes.

6.2. WP4: Amcanion

Roedd y gwaith a wnaed yn WP4 yn ymateb i dri amcan fel a ganlyn:

- dylunio a chynhyrchu pecyn cymorth pedagogaidd ar-lein sy'n seiliedig ar y Gymraeg (fel a ddisgrifir uchod) sy'n gweithio'n uniongyrchol â data'r corpws er mwyn cefnogi dysgu ac addysgu ieithyddol
- cynhyrchu rhestrau o eiriau addysgegol sy'n seiliedig ar eu hamlder

Un o nodweddion arloesol allweddol prosiect CorCenCC yw ei fod yn integreiddio corpws â phecyn cymorth addysgeg ar-lein. Mae Y Tiwtiadur yn gweithio'n uniongyrchol â data'r corpws er mwyn cefnogi dysgu ac addysgu ieithyddol drwy ddarparu cyfleoedd estynedig i archwilio darnau o Gymraeg go iawn. Lle ceir offerynnau corpws addysgegol cyn hyn sydd wedi bod yn ategiadau eilaidd i gorpysau a oedd eisoes yn bodoli, yn yr achos hwn, ac o'r cychwyn, roedd cynllun CorCenCC yn cynnwys rhyngwyneb addysgol er mwyn cefnogi dysgu ac addysgu'r Gymraeg. Ysbrydolwyd Y Tiwtiadur gan adnodd ar-lein y Compleat Lexical Tutor (Lextutor – https://www.lextutor.ca/ – Cobb, 2000) – un o'r pecynnau cymorth dysgu iaith ar-lein sydd wedi'i yrru gan ddata sydd â'r proffil uchaf ac sydd wedi'i gyrchu fwyaf (tua 15,000 o ddefnyddwyr y dydd). Mae'r bedair dasg a geir yn Y Tiwtiadur yn seiliedig ar rai o'r ymarferion mwyaf poblogaidd sydd i'w gweld yn Lextutor.

Un o'r nodweddion arloesol eraill yw'r ffaith fod y pecyn cymorth yn seiliedig ar ddysgu a yrrir gan ddata (Johns, 1991), lle 'mae dysgwyr yn archwilio'r dystiolaeth ac yn chwilio am batrymau yn y data y gallant gyffredinoli ohonynt' (Thompson, 2005: 10). Diben y pedwar ymarfer addysgegol yw galluogi ac annog dysgwyr (yn rhai iaith gyntaf ac ail iaith) i arsylwi ar batrymau iaith y Gymraeg, a dod i gasgliadau ohonynt er mwyn hwyluso dysgu. Mae hyn yn hanfodol bwysig i annibyniaeth dysgwyr (gweler Aston, 2001; Little, 2007). Yn hytrach na fo'r athro'n dweud wrth y myfyriwr sut mae'r iaith yn gweithio, caiff y myfyrwyr eu cefnogi wrth ei weithio allan drostynt eu hunain, gan ddefnyddio'r cysyniad o ddysgu anwythol (neu 'darganfodol'), a eiriolir yn glir gan y dull lluniadaethol o ddysgu. Ceir gwrthgyferbyniad rhwng y dull hwn a dysgu diddwythol (neu 'dan gyfarwyddyd tiwtoriaid'), fel pan ddarperir rheolau strwythurol i ddysgwyr. Mae integreiddio adnoddau dysgu i mewn i'r corpws, ac adeiladu'r corpws yn ôl anghenion dysgwyr, yn ei gwneud yn bosibl i ddysgwyr gyrchu data corpws perthnasol. Mae'r pedwar ymarfer addysgegol sydd wedi'u hymgorffori yn CorCenCC yn hwyluso'r broses o ganolbwyntio ar ffurf, y'i cydnabyddir fel elfen allweddol o ddysgu iaith yn effeithiol.

O gofio mai CorCenCC yw'r cyntaf o'i fath i dynnu ar amrywiaeth eang o genres ieithyddol, ffurfiau ieithyddol a ffyrdd ieithyddol o gynrychioli'r Gymraeg, roedd galluogi dysgwyr ac athrawon/tiwtoriaid i hidlo data'r corpws yn ôl ei fath a'i swm wrth gynhyrchu'r allbynnau a geir o ganlyniad yn drydedd nodwedd arloesol a ymgorfforwyd yn Y Tiwtiadur. Fel bod modd cymhwyso'r cyfleuster dysgu i bob oedran, un o agweddau allweddol y drydedd nodwedd arloesol hon yw'r gallu i hidlo darnau o destun sy'n debygol o gynnwys cynnwys amhriodol (fel rhegfeydd) fel nad ydynt yn ymddangos. Un o'r nodweddion y gellir ei gymhwyso i ddysgwyr o bob oedran yw'r gallu i wahaniaethu rhwng enghreifftiau o iaith o fathau gwahanol o ddata wrth roi ystyriaeth benodol i strwythurau cymhleth y mae modd iddynt amrywio rhwng siaradwyr (e.e. treiglo'r Gymraeg, neu rywfaint o ffurfiau enw lluosog, mewn darnau o destun sy'n amrywio o safbwynt eu ffurfioldeb). Yn y modd hwn, gall athrawon a dysgwyr reoli'r posibilrwydd am ddryswch mewn achosion o ddefnyddio mwy nag un ffurf.

6.3. WP4: Cyflawniadau

Roedd gwaith datblygu Y Tiwtiadur wedi'i lywio gan ddefnyddwyr o'r cychwyn. Ymgynghorwyd â thiwtoriaid/athrawon a dysgwyr, a oedd yn cynrychioli amrediad eang o lefelau cymhwysedd yn y Gymraeg ar draws y sectorau addysg ieithyddol gwahanol, ar amserau gwahanol er mwyn sicrhau y gellid dylunio, profi a gwella Y Tiwtiadur mewn modd ailadroddus. Dilynwyd nodau ac amcanion WP4 ar draws tri phrif gyfnod: (i) cyfnod ymgynghori, (ii) cyfnod datblygu cynnyrch, a (iii) cyfnod arddangos.

(i) Y cyfnod ymgynghori

Yn ystod y cyfnod ymgynghori, cafodd holiadur ei beilota gydag ychydig o ymarferwyr ym maes addysgu'r Gymraeg er mwyn archwilio pa adnoddau yr oedd dysgwyr ac athrawon yn eu defnyddio yn barod a pha rai y byddent am eu gweld yn cael eu datblygu yn ddelfrydol fel rhan o Y Tiwtiadur. Datblygwyd holiadur mwy manwl, a oedd wedi'i dargedu ar gyfer cynulleidfa fwy o faint, o ganlyniad i'w hadborth. Cafodd hwn ei ddosbarthu i athrawon a thiwtoriaid y Gymraeg yn ystod dwy gynhadledd genedlaethol, a chafodd ei rannu ar ffurf ar-lein yn ddiweddarach. Cafodd cyfanswm o 44 holiadur eu dychwelyd gan athrawon, hyfforddwyr, darlithwyr a thiwtoriaid y Gymraeg o amrediad eang o gyd-destunau – o'r bobl hynny sy'n addysgu mewn ysgolion cynradd (cyfrwng Cymraeg a chyfrwng Saesneg) i'r rheiny sy'n addysgu oedolion – a chynhaliwyd deg grŵp ffocws, gan gasglu safbwyntiau 55 o athrawon/tiwtoriaid ac 14 o ddysgwyr Cymraeg i Oedolion.

Yn ogystal â'r holiadur, gwnaethom gyfarfod wyneb yn wyneb ag athrawon a thiwtoriaid, er mwyn codi ymwybyddiaeth o'r corpws ac Y Tiwtiadur, ac i barhau i gasglu safbwyntiau y byddai'n helpu i lywio gwaith datblygu'r pecyn cymorth. Gwnaeth ymatebion yr holiadur, ynghyd â chyfarfodydd grŵp ffocws dilynol, ein helpu i nodi blaenoriaethau ar gyfer Y Tiwtiadur. Roedd y trafodaethau a gynhaliwyd gennym yn hynod ddefnyddiol, a gwnaeth y broses o gyfnewid syniadau a ddigwyddodd dros ystod y grwpiau ffocws, a thrwy'r adborth a gafwyd drwy'r holiaduron, wella'n dull o feddwl ynghylch sut i lywio'r gwaith wrth i ni symud ymlaen.

(ii) Y cyfnod datblygu cynnyrch a (iii) y cyfnod arddangos

Yn dilyn yr ymgynghoriad, cydweithredodd timau WP4 a WP5 i ddatblygu prototeipiau o'r offeryn. Yna, cafodd y gwaith hwn ei ddatblygu ymhellach gan J. Davies (a gafodd ei oruchwylio gan Teahan) mewn cydweithrediad ag Anthony. Cafwyd cyfle i arddangos y gwaith yr oedd wedi'i gwblhau ar Y Tiwtiadur hyd at y pwynt hwnnw yn ystod cynhadledd flynyddol y Ganolfan Dysgu Cymraeg Genedlaethol yn 2019, a chafwyd adborth cefnogol ac adeiladol gan y tiwtoriaid a fynychodd y gweithdai, gan gynnwys llawer o syniadau defnyddiol mewn perthynas â'i ddefnyddioldeb. Llywiodd yr adborth a gafwyd ganddynt weddill y gwaith datblygu. Roedd mewnwelediadau cadarnhaol yn cynnwys sut y gellid defnyddio Y Tiwtiadur gyda dysgwyr. Daeth tair prif thema i'r amlwg:

1. Cefnogi dealltwriaeth o safbwynt treiglo ar lefel frawddegol yn erbyn lefel eiriadurol

Un o nodweddion heriol y Gymraeg (a'r ieithoedd Celtaidd eraill) yw treiglo – proses forffoffonolegol lle mae newid ffonolegol yn cael ei achosi mewn set gaeedig o gytseiniaid

cyntaf geiriau pan fo geiriau penodol yn ymddangos mewn cyd-destunau cystrawennol penodol (Ball and Müller, 1992; Thomas and Mayr, 2010). Er enghraifft, mae sain y llythyren gyntaf 'c' /k/ yn ***c****ath* /kɑθ/ yn mynd trwy broses o feddaliad lle caiff /k/ ei meddalu i /g/ – /gɑθ/. Mae'r newid hwn yn digwydd ar ôl y fannod, *y*, ac ar ôl y rhif *dau* (gwrywaidd) / *dwy* (benywaidd), er enghraifft. Adlewyrchir y newid ffonolegol hwn yn y ffurf ysgrifenedig, lle ysgrifennir *cath* fel *gath*. Mewn rhai achosion, gellir trawsnewid cytsain gyntaf geiriau mewn tair ffordd, gan ddibynnu ar y cyd-destun (e.e. *cath* /kɑθ/ (ffurf waelodol), *gath* /gɑθ/ (treiglad meddal), *nghath* /ŋ̊ɑθ/ (treiglad trwynol) a *chath* /χɑθ/ (treiglad llaes). Gall y ffurfiau newidiol hyn beri problemau wrth ddarllen ac ysgrifennu, a gallant gael effaith ar ddatblygu/dysgu o safbwynt geirfa a llythrennedd, oherwydd eu bod yn ei gwneud yn anodd adnabod geiriau ac oherwydd bod y rheolau ar gyfer y treigladau yn gymhleth weithiau. Gall enghreifftiau a yrrir gan gorpysau dynnu sylw i ffurf a helpu dysgwyr i nodi lle, pryd a sut mae geiriau'n newid ar draws cyd-destunau a faint o amrywiaeth a geir o safbwynt mynegi ffurfiau 'targed'.

2. Llenwi bylchau mewn cymorth geiriadurol

Mewn perthynas uniongyrchol â threiglo, yn ogystal â ffurfiau newidiol eraill, mae llawer o ddysgwyr yn methu â dod o hyd i eiriau mewn geiriaduron papur a geiriaduron ar-lein oherwydd eu bod yn chwilio am y ffurf dreigledig (e.e. ***g****ath*) ac nid y ffurf waelodol (h.y. ***c****ath*), neu oherwydd eu bod wedi camsillafu neu gam-gynrychioli'r gair mewn testun. Bydd y corpws yn galluogi defnyddwyr i ddarganfod sut y mae ffurf sydd o ddiddordeb iddynt yn cael ei defnyddio fel arfer mewn mathau gwahanol o genres a chyfryngau.

3. Cefnogi dysgwyr wrth iddynt werthuso'u gwaith ysgrifenedig

Un o nodweddion allweddol Y Tiwtiadur yw'r opsiwn sydd ar gael i ddysgwyr a/neu athrawon/tiwtoriaid i ddefnyddio'r feddalwedd i godio darn o destun o'u dewis o ran pa mor anodd ydyw yn nhermau amlder y ffurfiau a ddefnyddir. Gall hyn fod yn ddefnyddiol wrth benderfynu pa mor addas yw darn o destun ar gyfer dysgwr neu grŵp o ddysgwyr. Yn ogystal, gall dysgwyr broffilio'u gwaith ysgrifenedig eu hunain, er mwyn gwerthuso pa mor eang yw eu gwybodaeth am eirfa a'u defnydd ohoni.

Yn ogystal â'r mewnwelediadau hyn o ran sut y gellid rhoi'r offerynnau hyn ar waith, cafwyd myfyriadau gwerthfawr gan randdeiliaid ar eu pryderon ynghylch defnyddio adnodd o'r fath yn y dosbarth. Codwyd pum prif broblem, fel a amglygir isod:

1. *Modelu iaith 'anghywir'*

Un o'r pryderon a gododd sawl tro ymhlith athrawon a thiwtoriaid o bob cyd-destun addysgol oedd y gallai data'r corpws, oni iddo gael ei 'gywiro', fodelu'r union ffurfiau ac ymadroddion y mae athrawon yn ceisio cael gwared ohonynt o ymdrechion eu dysgwyr. Gan nad yw corpws yn gwahaniaethu'n benodol rhwng yr hyn a ystyrir yn gywir neu'n anghywir ar lefel ragnodol (ar wahân i'r hyn a wneir yn ôl amlder neu ymddangosiad geiriau), cynghorwyd athrawon i ymgyfarwyddo â'r darnau o destun y byddent yn eu defnyddio yn ystod sesiynau ymlaen llaw, ac i nodi unrhyw beth y byddent am gynghori'r dysgwyr yn ei gylch. Gwnaeth y dull hwn gydnabod bod y berthynas rhwng dulliau disgrifiadol a rhagnodol o addysgu iaith

yn gymhleth. Roedd yn cydnabod bod dyletswydd ar athrawon i hysbysu dysgwyr am y ffurfiau y bydd pobl eraill yn eu hystyried yn rhai 'anghywir' (gan fod ymwybyddiaeth o'r fath yn un o'r agweddau sydd ynghlwm wrth wybodaeth am yr iaith). Ar yr un pryd, byddai'r dull â llaw yn annog athrawon i gwestiynu eu credoau eu hunain ynghylch yr hyn sy'n 'dderbyniol' fel ffurf darged, neu beidio, o gofio'r defnydd a dystir.

2. *Iaith sarhaus*

Cododd llawer o athrawon (o'r sector addysg gynradd yn enwedig) bryderon am y posibilrwydd o gyrchu cynnwys amhriodol ar ddamwain, gan ganolbwyntio'n bennaf ar gynnwys a oedd yn ymwneud â geiriau amhriodol neu sarhaus (fel rhegfeydd). Er bod gwaith codio ar gyfer lefelau a mathau o fod yn sarhaus y tu hwnt i gwmpas y prosiect presennol (na fyddai, mewn rhai achosion, i'w cael ar lefel gair unigol bob amser), tagwyd y corpws am eiriau rhegi a chynnwys o natur sensitif benodol ar y lefel destun. Er bod y cyfrifiadau ystadegol y mae'r offerynnau addysgeg yn seiliedig arnynt yn cyfeirio at y 11 miliwn o eiriau sydd wedi'u cynnwys yn CorCenCC, mae pedwar ymarfer y pecyn cymorth pedagogaidd yn hidlo'r darnau testun hyn fel nad ydynt yn ymddangos.

3. *Cymhlethdod y rhyngwyneb*

Un o nodau clir CorCenCC oedd datblygu corpws a fyddai'n hwylus i ddefnyddwyr ac y byddai'n gallu cael ei drosglwyddo i faes addysg yn hawdd. Er mwyn i hynny weithio, roedd yn rhaid i'r rhaglen fod yn addas i'r diben ac yn hwylus ei ddefnyddio. Cododd yr ymdeimlad hwn ymhlith rhai o'r tiwtoriaid a'r athrawon y gwnaethom gwrdd â nhw. Helpodd y safbwyntiau hyn i sicrhau bod rhyngwyneb syml a chlir yn perthyn i'r pedwar ymarfer addysgegol a oedd yn reddfol o safbwynt y defnyddiwr. Ar yr un pryd, roedd y rhyngwyneb hwn yn adlewyrchu'r rhyngwyneb a ddefnyddir gan CorCenCC er mwyn sicrhau cydlyniant ar draws y ddau ac i ennyn hyder mewn athrawon o safbwynt dilyniant o ddefnyddio'r offerynnau addysgeg i ddefnyddio'r corpws yn ehangach pe byddent am wneud hynny.

4. Hygyrchedd

Er bod gwaith datblygu platfformau amgen ar gyfer cynnal y corpws y tu hwnt i'r hyn sydd dan sylw yn yr astudiaeth bresennol, roedd yn glir, mewn ysgolion yn enwedig, lle ceir mynediad cyfyngedig i gyfrifiaduron yn aml, y byddai ap y gellid ei lawrlwytho ar ffonau yn ddefnyddiol. Mae ysgolion yn defnyddio apiau sydd eisoes yn bodoli ar gyfer y Gymraeg yn barod, fel Duolingo, ac maent yn gweld bod y platfform hwnnw'n ddefnyddiol, felly gofynnodd athrawon a fyddai'n bosibl datblygu platfform tebyg ar gyfer y prosiect presennol. Ar y cyfan, roedd yn amlwg y bydd datblygu ap ar gyfer ei ddefnyddio yn y dosbarth yn addasiad gwerthfawr o Y Tiwtiadur yn y dyfodol. Gan nad oedd hwn yn un o nodau'r prosiect presennol, ac y byddai'n cymryd gwaith ymchwil ac adnoddau sylweddol i'w gwblhau, mae wedi'i nodi fel ystyriaeth bwysig ar gyfer gwaith dilynol.

5. Nifer brawychus yr enghreifftiau yn yr allbwn

Mae CorCenCC yn gorpws o dros 11 miliwn o eiriau y mae modd ei chwilio. Mae hyn yn golygu y bydd rhai chwiliadau'n arwain at nifer anferth o allbynnau. I'r bobl hynny nad oes

ganddynt brofiad o ddefnyddio corpysau a'r mathau o allbynnau a gynhyrchir ganddynt, gall swm yr allbynnau a gynhyrchir fod yn llethol, a gallai atal dysgwyr ac athrawon/tiwtoriaid. Am y rheswm hwnnw, un o elfennau arloesol Y Tiwtiadur yw ei fod yn rhoi'r gallu i athrawon, tiwtoriaid a dysgwyr reoli swm yr allbwn a gynhyrchir gan ymholiadau (fel a eiriolir gan y dull dysgu a yrrir gan ddata). Yn ogystal â'r gallu i ddethol testunau neu gategorïau semantig wrth archwilio darnau o destun, mae'r offeryn yn rhoi'r gallu i athrawon/tiwtoriaid a dysgwyr gyfyngu ar hyd y darnau o destun wrth orchuddio geiriau yn yr ymarfer Llenwi Bylchau, ac yn gosod uchafswm o 20 o achosion ar yr allbwn a gynhyrchir ar gyfer yr ymarferion Nodi Geiriau a Geiriau yn eu Cyd-destun ac ati. Gyda'i gilydd, mae'r nodweddion hyn yn cynyddu lefel annibyniaeth y dysgwr a'r athro/tiwtor fel eu bod yn gallu sicrhau bod y pecyn cymorth yn gweithio iddynt.

6.4. WP4: Cyfraniadau allweddol

Drwy'r gwaith o ddatblygu rhyngwyneb addysgegol, a arweinir gan y cysyniad o ddysgu ac asesu a yrrir gan ddata, mae WP4 wedi cyfrannu (i) adnodd addysgegol newydd sydd (ii) wedi'i dynnu o gorpws ar-lein o Gymraeg cyfoes, o fath (iii) nad yw wedi bodoli erioed o'r blaen ar gyfer addysgu'r Gymraeg, a (iv) y gall fod yn fodel ar gyfer gwaith tebyg o safbwynt ieithoedd lleiafrifol eraill. Mae wedi gwneud cyfraniad gwerthfawr i faes dysgu ac addysgu ieithoedd a, gan ei fod o natur ffynhonnell agored, mae ar gael i gefnogi dysgwyr fel rhan o'u dysgu anwythol, ni waeth eu hoedran, eu lefel gallu a'u lleoliad daearyddol. Mae'r adnodd yn cynnig cyfle newydd ac unigryw ar gyfer ysgolion yng Nghymru i ymgymryd â'r cysyniad o ddysgu a yrrir gan ddata ac i ymgysylltu â'u dulliau addysgeg eu hunain a arweinir gan y corpws, a'u datblygu. Yn ogystal â hyn, unwaith y bydd athrawon a dysgwyr wedi'u cyflwyno i'r corpws drwy Y Tiwtiadur, gallent hefyd deimlo'n ddigon hyderus i archwilio'r corpws mewn ffyrdd eraill drwy ddefnyddio prif offerynnau ymholi CorCenCC.

Cafwyd galwadau amrywiol dros y blynyddoedd diwethaf am gorpws sy'n gallu llywio'r dull o ddarparu'r Gymraeg (NFER, 2008: 48; Llywodraeth Cymru 2013: 27, 71; Mac Giolla Chríost et al., 2012). Fel corpws cyfoes o'r Gymraeg sy'n cynnwys pecyn cymorth pedagogaidd integredig (Y Tiwtiadur), mae CorCenCC yn diwallu'r angen hwnnw, gan lywio gwaith ysgrifenedig y cwricwlwm, gwaith asesu iaith ac adnoddau dysgu iaith yn yr un modd ag y mae corpysau tebyg yn ei wneud ar gyfer y Saesneg (e.e. mae'r Cambridge English Corpus (CEC) yn llywio adnoddau addysgu Cambridge English Language; mae'r British National Corpus (BNC) yn llywio adnoddau Pearson Longman). Gallai datblygu set gyfatebol lawn ac annibynnol o adnoddau addysgu'r Gymraeg, sy'n seiliedig ar CorCenCC, fod yn drywydd posibl i'r gwaith hwn yn y dyfodol.

Yn unol â'r disgwyliadau a bennir yn y *Cwricwlwm i Gymru: 2022* newydd, bydd adnodd o'r fath yn helpu i ddatblygu galluoedd ymwybyddiaeth ieithyddol dysgwyr a gloywi eu sgiliau Cymraeg mewn ffordd naturiolaidd, a fydd, yn y pen draw, yn cael effaith ar agenda *Cymraeg 2050: Miliwn o Siaradwyr* Llywodraeth Cymru (Llywodraeth Cymru, 2017), sydd â'r uchelgais o sicrhau miliwn o siaradwyr Cymraeg erbyn 2050.

6.5. WP4: Cymwysiadau ac effaith

Mae defnyddio data o gorpysau i gefnogi dysgu ieithoedd mewn ysgolion yn arfer sy'n datblygu'n gyflym, ond nid yw'r dull o'i weithredu'n effeithiol wedi'i ddatblygu'n llawn a phrin yw'r gwaith ymchwil i'w effeithiolrwydd. Yn y Cwricwlwm i Gymru: 2022 newydd, bydd yn ofynnol i blant ddysgu am gysyniadau ieithoedd, dadansoddi mân-wahaniaethau ieithyddol, a deall sut maent yn wahanol mewn ieithoedd gwahanol. Mae data o gorpysau wedi'i alinio'n berffaith â'r gwaith o hyrwyddo sgiliau a gwybodaeth feta-ieithyddol, yn enwedig mewn cyd-destun dwyieithog o'r math a geir yng Nghymru. Cam pwysig nesaf fydd rhannu'r adnodd ar-lein, rhad ac am ddim hwn i ysgolion yng Nghymru. Bydd hyn yn golygu gweithio'n agos â CBAC, dylunwyr cwricwlwm a Llywodraeth Cymru wrth nodi'r arfer gorau ar gyfer defnyddio CorCenCC mewn cyd-destunau dysgu ac addysgu ledled Cymru, a modelu'r dull o'i weithredu mewn cyd-destunau iaith leiafrifol eraill mewn lleoedd eraill. Gallai'r gwaith hwn arwain at brosiectau ymchwil sydd wedi'u hariannu ar gyfer gwerthuso effeithiolrwydd cymwysiadau amrywiol CorCenCC ac Y Tiwtiadur gyda mathau gwahanol o ddysgwyr gyda'r bwriad o ymestyn ei ymarferoldeb lle bo hynny'n briodol. Er enghraifft, gellid addasu ymarferion ar gyfer ap ffôn neu blatfformau technolegol eraill, gan gynyddu defnydd yr adnodd a chynyddu ei effaith ar faes addysg.

7. Pecyn Gwaith 5: Adeiladu'r seilwaith i letya CorCenCC

7.1. WP5: Disgrifiad

Roedd WP5 yn ymwneud â'r elfennau technegol o adeiladu CorCenCC, gan greu offerynnau i gefnogi pob cam o'r broses o adeiladu'r corpws, o gasglu data (gan ganolbwyntio ar yr ap cyfrannu torfol), trwy'r cam o goladu (drwy ddefnyddio'r offerynnau rheoli data), i'r gwaith ymholi a dadansoddi (y rhyngwyneb ar y we).

Spasić a arweiniodd WP5, gan weithio gyda Knight, Rayson a Piao, a Neale a Muralidaran, sef dau o gynorthwywyr ymchwil y prosiect. Darparodd Anthony (ieithydd corpws, arbenigwr technoleg addysgol a chrëwr Antconc), Scannell (gwyddonydd cyfrifiadurol ag arbenigedd mewn prosesu iaith naturiol (NLP), cyfieithu peirianyddol ac ieithoedd lleiafrifol) a Donnelly (ieithydd cyfrifiadurol a weithiodd yn agos ar ddatblygu corpysau Cymraeg blaenorol) arbenigedd technegol ac ymgynghorol ychwanegol.

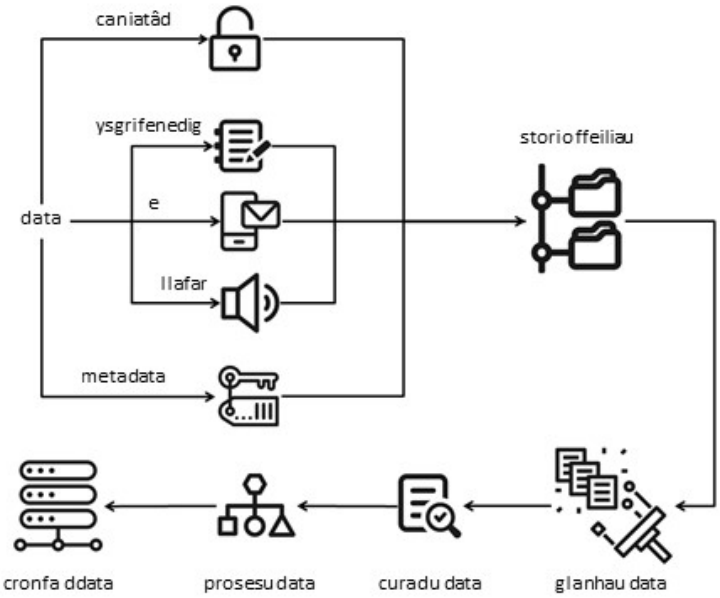
7.2. WP5: Amcanion

Bwriad WP5 oedd datblygu seilwaith cyfrifiadurol i gefnogi'r gwaith o gasglu a storio'r swm mawr hwn o destun a data dadansoddol mewn modd systematig, ynghyd â rhyngwyneb hwylus ei ddefnyddio er mwyn ei gwneud yn bosibl rhyngweithio â'r data hyn ar-lein. Un o elfennau pwysig y gwaith hwn oedd cynllunio ac adeiladu system ystorio a fyddai'n ei gwneud yn bosibl ychwanegu data newydd i'r corpws dros amser, fel y gallai defnyddwyr gefnogi'r gwaith o gynnal y corpws ac fel y byddai cyfraniadau i'r corpws yn fenter gymdeithasol. Cafodd cyfres o offerynnau dadansoddi'r corpws ei datblygu ar ben yr ystorfa er mwyn cefnogi swyddogaethau sydd fel arfer wedi'u hintegreiddio i mewn i gorpysau cyfoes, fel peiriannau cydgordio ac offerynnau cydleoli Gair Allweddol mewn Cyd-destun

(KWIC), offerynnau chwilio a threfnu, rhestrau amlder geiriau, peiriannau dadansoddi geiriau allweddol a chyfleusterau profi ystadegol.

Mae Ffigur 2 yn dangos y llif gwaith ar gyfer casglu data. Amlygir y gwahaniaeth rhwng y tri phrif fath o iaith (llafar, ysgrifenedig ac electronig (e)), gan fod angen cwblhau dulliau prosesu gwahanol arnynt cyn y gallai'r data gael ei integreiddio i mewn i'r corpws. Fel y mae Ffigur 2 yn dangos, cafodd yr holl wybodaeth berthnasol am gyfranogwyr a metadata disgrifiadol ei gofnodi ar yr un pryd ag y cafodd y data ei gasglu. Roedd cael caniatadau i rannu'r data mewn adnodd cyhoeddus ar-lein yn hanfodol i waith datblygu CorCenCC. Cafwyd y caniatadau hyn gan yr endidau cyfreithiol perthnasol (e.e. perchennog yr hawlfraint; y siaradwr ei hun) cyn y cafodd y data ei gasglu a'i storio'n lleol. Cafodd y data crai ynghyd â'r caniatadau a metadata cyfatebol eu hadneuo mewn system storio ffeiliau leol. Yn dilyn hynny, cafodd fformatau data gwahanol eu safoni'n destun plaen. Mae'n bosib prosesu testun plaen yn awtomatig gan ddefnyddio offerynnau prosesu iaith naturiol (NLP), pe byddai angen ychwanegu haen arall o fetadata ieithyddol yn ddiweddarach; mae hyn yn helpu i ddiogelu'r corpws at y dyfodol, drwy ei gwneud yn bosibl ychwanegu gwybodaeth ychwanegol.

Ffigur 2 Llif gwaith ar gyfer casglu data yn CorCenCC



7.3. WP5: Cyflawniadau

Un o nodweddion arloesol allweddol prosiect CorCenCC oedd ailddiffinio'r ffyrdd o gynllunio ac adeiladu corpysau ieithyddol, gan alinio'r dulliau ag oes Gwe 2.0 mewn modd mwy cryno. I'r perwyl hwn, cymerwyd camau i adeiladu a gwerthuso system sy'n ei gwneud yn bosibl casglu data llafar gan siaradwyr drwy gyfrannu torfol. Mae cyfrannu torfol yn ffordd o gasglu adnoddau (enghreifftiau ieithyddol yn yr achos hwn) gan y cyhoedd, drwy ofyn i wirfoddolwyr gymryd rhan. Gwnaethom hwyluso cyfrannu torfol drwy ddefnyddio ap y mae modd ei gynnal ar unrhyw ddyfais sydd â chysylltiad â'r we; gweithiodd ar ddwy ffurf: fel cymhwysiad ffôn ac fel gwefan ryngweithiol. Er mwyn cynyddu'r sylfaen o ddefnyddwyr posibl, gweithredwyd fersiwn symudol yr ap ar blatfformau iOS (h.y. Apple) ac Android.

Roedd y cymhwysiad yn sicrhau bod y profiad o gyfrannu at y corpws yn un personol iawn, gan roi perchenogaeth i ddefnyddwyr dros eu recordiadau eu hunain, a rheolaeth drostynt.

Cafodd yr holl ddata crai (yn llafar, yn ysgrifenedig ac yn electronig) ei storio mewn modd systematig o fewn strwythur ffolder a oedd wedi'i ddiffinio ymlaen llaw, ac a gyfatebodd â'r fframwaith samplu. O hynny ymlaen, aeth y data trwy'r prosesau glanhau a churadu perthnasol. I gefnogi mynediad cydweithredol ar gyfer defnyddwyr lluosog gan aelodau'r tîm (o ymchwilwyr i drawsgrifwyr) ar draws safleoedd gwahanol (ar draws ac o fewn y sefydliadau lluosog a oedd ynghlwm wrth y prosiect), datblygwyd offeryn rheoli data ar-lein ar ben y system storio ffeiliau. Darparodd ryngwyneb defnyddiwr graffigol (GUI) a hwylusodd y broses o lanlwytho data crai, mynegeio'r metadata cyfatebol a chofnodi trawsnewidiadau data dilynol, gan sicrhau y gallai pob ymchwilydd fonitro cynnydd holl elfennau'r broses o adeiladu'r corpws yn fanwl.

Unwaith yr oedd y darnau o destun wedi'u trosi i fformat testun plaen, cawsant eu nodi â haenau o fetadata sosioieithyddol (e.e. ffynhonnell, genre, tarddiad daearyddol) a fyddai'n cael eu defnyddio i ymholi'r data, a chawsant eu tagio'n awtomatig. Fel a ddisgrifiwyd yn Adran 4, datblygodd tîm WP2 CorCenCC CyTag (Neale et al., 2018) er mwyn cyflawni'r broses o bennu tagiau. Mae CyTag yn gyfres o offerynnau prosesu iaith naturiol ar lefel wyneb ar gyfer y Gymraeg sy'n seiliedig ar gysyniad gramadeg cyfyngiadol (Karlsson, 1990, Karlsson et al., 1995). Mae'n cefnogi'r syniad o segmentu testun, gan gynnwys hollti brawddegau a thocyneiddio, yn ogystal â phennu tagiau i rannau ymadrodd a'u lemateiddio. Mae'n cynnig datrysiad pwrpasol ar gyfer rhagbrosesu'r Gymraeg mewn modd ieithyddol sylfaenol, gan gynnwys set dagiau sydd â chyfoeth digonol o ddata ynddi i gipio mympwyon yr iaith – a geir ar ei ffurfiau llafar yn enwedig. I hwyluso'r broses o gyflawni dadansoddiad semantig ar ddata am y Gymraeg ar raddfa fawr, cafodd yr holl ddata a ragbroseswyd ei ddynodi ymhellach yn ôl categorïau semantig gan ddefnyddio'r CySemTagger a ddatblygwyd yn WP3 (gweler Adran 5).

Cafodd y corpws ei storio a'i reoli mewn cronfa ddata berthynol lle'r oedd yn bosibl cyrchu data mewn modd diogel gan ddefnyddwyr lluosog ac ar yr un pryd â'i gilydd. Er mwyn rhannu'r data ar-lein, gwnaethom ddatblygu rhyngwyneb ar y we ar gyfer y gronfa ddata. Y prif reswm dros ddatblygu rhyngwyneb pwrpasol yn hytrach nag ailddefnyddio datrysiad a oedd eisoes yn bodoli, fel CQPweb (Hardie, 2012) oedd y gofyniad i deilwra ei ymarferoldeb i fetadata penodol CorCenCC a'i ddarpar ddefnyddwyr. Er mwyn casglu gwybodaeth am ofynion defnyddwyr, gwnaethom ddefnyddio'r cyfryngau cymdeithasol i gynnal arolwg o ddefnyddwyr corpysau ar y pryd. Cafwyd ymatebion gan gyfanswm o 62 o unigolion, a gwnaeth eu mewnbwn nodi'r gofynion allweddol o safbwynt ymarferoldeb.

Un o'r ystyriaethau pwysig ar gyfer datblygu seilwaith y corpws oedd gwirio'i fod yn addas i'r pwrpas. Gwerthusodd grŵp o ieithyddion corpws ddefnyddioldeb ac ymarferoldeb y rhyngwyneb ar y we. Roedd y broses werthuso hon yn cynnwys cyfuniad o holiaduron ac ymarferion siarad yn uchel. Yn gyffredinol, roedd y cyfranogwyr yn meddwl bod y system yn ddefnyddiol yn nhermau diwallu eu hanghenion gwybodaeth o fewn cwmpas eu gweithgareddau proffesiynol. Roedd yr ymarferoldeb yn hawdd ei ddeall heb orfod cyfeirio at y sgrin gymorth. Cytunodd pob cyfranogwr y byddent yn debygol o fabwysiadu'r system a'i hargymell i ieithyddion eraill.

7.4. WP5: Cyfraniadau allweddol

Mae creu seilwaith corpws newydd yn golygu gwaith sylweddol. Mae'r mwyafrif o ymchwilwyr corpws yn defnyddio meddalwedd sydd eisoes yn bodoli i ddadansoddi'r darnau o destun y maent yn eu casglu. Fodd bynnag, yn yr achos hwn, nid oedd y fath hon o feddalwedd eisoes yn bodoli, ac felly roedd angen ei hadeiladu cyn y gallai'r darnau o destun roeddem wedi'u casglu gael eu mynegeio'n briodol er mwyn eu mewnbynnu i'r corpws. Felly, roeddem yn mynd i'r afael â sawl her sylweddol ar y cyd a oedd yn bwysig ar gyfer datblygu gwaith ymchwil corpysau ieithyddol .

Yn ail, roedd yn rhaid i ni ddylunio llawer o'r seilwaith cyfrifiadurol gwaelodol ar gyfer pennu tagiau i'r iaith, a'i dadansoddi, gan ddechrau o'r dechrau, gan gofio bod llawer o nodweddion yn perthyn i'r Gymraeg, gan gynnwys gwahaniaethau gramadegol, nad yw'n hawdd eu trosglwyddo o ieithoedd sydd ag adnodd corpws sylweddol sydd eisoes yn bodoli (yn enwedig Saesneg), ac amrywiaeth ranbarthol a chyweiriol sylweddol sy'n deillio o hanes cymdeithasol arbennig yr iaith. Mae pileri allweddol y seilwaith yn cynnwys fframwaith sy'n cefnogi'r broses o gasglu metadata, ap symudol arloesol sydd wedi'i ddylunio ar gyfer casglu data llafar (gan ddefnyddio dull cyfrannu torfol), cronfa ddata o'r ochr gefn sy'n storio data curedig, a rhyngwyneb ar y we sy'n galluogi defnyddwyr i ymholi'r data ar-lein. Drwy ddefnyddio tagiau Cymraeg, rydym wedi sicrhau nad yw'r corpws yn cael ei ganfod yn offeryn allanol (Saesneg) sydd wedi'i arosod ar y Gymraeg, ac na fydd yn bosib ei ganfod felly, ond yn hytrach ei fod yn perthyn i Gymru a'r Gymraeg. Yn y modd hwn, anogir defnyddwyr i lwyr gefnogi'r iaith, nid yn unig fel ffynhonnell wybodaeth ond hefyd fel y cyfrwng y gellir ei hastudio drwyddo. Ar yr un pryd, bydd y ffaith fod rhyngwyneb Saesneg ychwanegol ar gael yn sicrhau pwynt mynediad ar gyfer y nifer fawr o bobl hynny sydd â diddordeb yn y Gymraeg sy'n fwy na'u hyfedredd ynddi, gan gynnwys y miloedd o ddysgwyr Gymraeg.

Yn drydydd, rydym wedi creu offerynnau sydd ar gael yn rhad ac am ddim er mwyn i bobl eraill eu haddasu wrth greu eu corpysau eu hunain. Rydym yn ymrwymedig yn enwedig i gefnogi'r gwaith o adeiladu corpysau ar gyfer ieithoedd lleiafrifol eraill, ac mae ein model a yrrir gan ddefnyddwyr yn llywio prosiectau o'r fath yn uniongyrchol trwy ddarparu templed ar gyfer datblygu corpysau mewn unrhyw iaith arall.

Yn bedwerydd, drwy gasglu tîm rhyngwladol o arbenigwyr ynghyd, rydym wedi gallu rhoi'r datblygiadau arloesol technolegol diweddaraf ar waith, a datblygu ein syniadau newydd ein hunain, gan fwrw golau ar drywydd ar gyfer gwaith yn y dyfodol yn oes Gwe 2.0. Mae'r ap cyfrannu torfol yn un o'r cyntaf o'i fath i gael ei ddefnyddio ar gyfer adeiladu corpws cytbwys o ddata iaith naturiol trwy ategu dulliau mwy traddodiadol o gasglu data, ac sydd wedi mynd i'r afael yn llwyddiannus â phroblem sylweddol a chyson o safbwynt casglu data llafar o ansawdd uchel â chaniatâd. Yn ogystal, rydym wedi dangos ei bod yn bosibl darganfod yr adnoddau dynol angenrheidiol ar gyfer trawsgrifio data o'r fath ac ar gyfer cwblhau'r gwaith angenrheidiol o bennu tagiau â llaw, hyd yn oed ar gyfer iaith sydd â charfan gymharol fach o siaradwyr rhugl.

7.5. WP5: Cymwysiadau ac effaith

Er y datblygwyd y seilwaith cyfrifiadurol ar gyfer casglu gwybodaeth am y Gymraeg, mae'n bosib i'w gynllun gael ei ailddefnyddio i gefnogi gwaith datblygu corpysau yng nghyd-destunau ieithoedd lleiafrifol neu brif ieithoedd eraill, ac mae hynny'n ehangu defnyddioldeb ac effaith y gwaith hwn.

8. Crynodeb o gymwysiadau posibl a'u heffaith

Fel y disgrifiwyd yn gynharach yn y ddogfen hon, mae gan bob rhan o'r prosiect (fel y nodweddir gan y pecynnau gwaith) gymwysiadau gwerthfawr sy'n cynnig manteision cymdeithasol, economaidd a/neu academaidd. Ar lefel gymdeithasol, mae'r corpws yn cynnig y cyfle i ddeall y Gymraeg fel iaith fyw sy'n cael ei defnyddio. Mewn termau economaidd, mae'r corpws yn cynnig lle i ddatblygu adnoddau newydd gwerthfawr ar gyfer dysgwyr a defnyddwyr Cymraeg, gan gynnwys y posibilrwydd o greu geiriadur sy'n seiliedig ar y corpws ac amrediad o offerynnau technolegol sydd wedi'u llywio gan ddata a allai gynnwys apiau dysgu iaith, cynhyrchu testun rhagfynegol, offerynnau prosesu geiriau, cyfieithu peirianyddol, ac offerynnau adnabod llais a chwilio'r we. Yn anuniongyrchol, bydd cefnogaeth y corpws ar gyfer sicrhau'r canlyniadau cymdeithasol ac economaidd hyn yn hyrwyddo cydnabyddiaeth y Gymraeg fel elfen sylweddol o dirwedd ieithyddol y DU a'r byd. Mae'r cymwysiadau academaidd posibl yn eang ac yn amrywiol, fel a ganlyn:

- Ieithyddiaeth corpws: Bydd cynllun arloesol CorCenCC yn llywio ac yn arwain y gwaith o ddatblygu corpysau yn y dyfodol a'u defnydd, mewn unrhyw iaith, drwy ddarparu adnoddau ffynhonnell agored a'r protocolau ar gyfer dulliau torfol o gasglu a dadansoddi data ac ar gyfer integreiddio swyddogaethau gwaith ymchwil a dysgu ac addysgu.
- Caffael iaith a dwyieithrwydd: Mae'r cyfleusterau dysgu a yrrir gan ddata sydd wedi'u hymgorffori yn cynnig cyfle unigryw ar gyfer ymchwilio i ymddygiad dysgwyr annibynnol a dysgu cyfunol. Mae CorCenCC yn ymgorffori cyfres o offerynnau sy'n seiliedig ar amlder ar gyfer gwneud gwaith ymchwil i batrymau o gaffael iaith ac i broffilio testun a dysgwyr.
- Sosioieithyddiaeth, tafodieitheg a dadansoddiadau morffogystrawennol: Mae CorCenCC yn ymestyn cwmpas y data sgyrsiol sydd ar gael o'r corpws Siarad (www.bangortalk.org.uk) er mwyn cynnwys siaradwyr o amrediad ehangach o ranbarthau daearyddol. Bydd hyn yn rhoi mwy o wybodaeth inni o safbwynt faint o gyswllt ieithyddol a geir rhwng y Gymraeg a'r Saesneg. Bydd yn hwyluso'r gwaith o archwilio amrywiaeth sosioieithyddol ym mhatrymau defnydd yr iaith o safbwynt geiriadurol, gramadegol, semantig ac ymarferol, ac o safbwynt priodweddau defnyddio'r iaith sy'n seiliedig ar acenion, o fewn ac ar draws rhanbarthau, ac yn ôl proffiliau defnyddwyr y Gymraeg, a thrwy hynny ceir dealltwriaeth ddyfnach o ddynameg gymdeithasol y Gymraeg fel iaith fyw ac sy'n adfywio.

- Cynllunio ieithyddol: Mae CorCenCC yn darparu data llinell sylfaen y gellir ei ddefnyddio i ymchwilio i ddefnydd yr iaith yng nghyd-destun polisïau sy'n ymwneud â defnyddio'r Gymraeg ym meysydd gweinyddiaeth gyhoeddus, masnach ac addysg.
- Geiriadura: Bydd CorCenCC yn uniongyrchol berthnasol i waith y tîm sydd wrthi'n diweddaru Geiriadur Prifysgol Cymru (sy'n bartner prosiect), drwy ddarparu tystiolaeth sylfaenol o ddefnydd geiriadurol o ffynonellau y gellir eu priodoli ac y gellir eu cynnwys yn niwygiadau'r geiriadur yn y dyfodol.
- Technoleg gyfrifiadurol a thechnoleg cyfieithu: Fel corpws â thagiau a bennir, bydd yn bosibl defnyddio CorCenCC ar gyfer datblygu systemau cyfieithu peirianyddol drwy gyfrannu at beiriant cyfieithu peirianyddol ystadegol, ac iddo hwyluso'r broses o haniaethu rheolau gramadegol o'r corpws. Bydd meysydd eraill o brosesu iaith naturiol (e.e. cywiro sillafu, rhagfynegi geiriau, meddalwedd gynorthwyol ar gyfer y Gymraeg) yn manteisio o'r modelau ieithyddol mwy manwl gywir, ac sydd ar gael yn ehangach, a gaiff eu cynhyrchu o ganlyniad i ddadansoddi data o CorCenCC.
- Meysydd ymchwil eraill: Mae CorCenCC yn cynnig cyfleoedd newydd i academyddion sy'n astudio meysydd arddulleg a llenyddiaeth, y cyfryngau, seicoieithyddiaeth, ymarferoleg, busnes, iechyd a meddygaeth, a seicoleg ymestyn eu gwaith ymchwil i gyd-destun y Gymraeg.

Mae CorCenCC yn adnodd sydd ar gael yn rhad ac am ddim o dan drwydded agored a fydd, o'i gyfuno â'i gynllun, a'r dull o'i adeiladu a yrrir gan ddefnyddwyr, yn cynyddu ei effaith gymdeithasol bosibl, gan lywio gwaith a gweithgareddau defnyddwyr presennol y Gymraeg, a rhai'r dyfodol, mewn sawl maes hanfodol bwysig. Mae meysydd ymarferwyr posib nad ydynt yn academaidd a meysydd proffesiynol posib yn cynnwys y canlynol:

- Dysgu ac addysgu ail iaith: Fel a drafodwyd yn Adran 6, mae adroddiadau ar addysgu Cymraeg i Oedolion (Mac Giolla Chríost ac eraill, 2012; Llywodraeth Cymru, 2013) wedi tynnu sylw at yr angen am gorpws o Gymraeg cyfoes fel ffordd o wella effeithiolrwydd mentrau dysgu'r Gymraeg. Mae CorCenCC yn gweithio i ddiwallu'r angen hwn. Drwy lywio gwaith ysgrifenedig y cwricwlwm, gwaith asesu'r iaith ac adnoddau dysgu'r iaith fel y mae corpysau tebyg yn ei wneud yn effeithiol yn y Saesneg (e.e. CEC a BNC), bydd CorCenCC yn hwyluso dysgu a yrrir gan ddata, gan wella effeithiolrwydd addysgu Cymraeg fel ail iaith (sy'n orfodol ymhob ysgol yng Nghymru hyd at ddiwedd Cyfnod Allweddol 4). Rhagwelir y bydd yr effeithiau yn y tymor canolig yn cynnwys gwelliant yn effeithiolrwydd dysgu'r iaith; gwell ymwybyddiaeth gan athrawon a dysgwyr, o safbwynt mân-wahaniaethau, o'r amrywiaeth gynhenid a naturiol a geir yn y Gymraeg fesul rhanbarth, genre, a math o siaradwr; cynnydd yn hyder siaradwyr Cymraeg ynghylch dilysrwydd eu patrymau defnydd eu hunain; a gwell ymwybyddiaeth o'r Gymraeg, a balchder ynddi, fel ffordd fyw a datblygol o fynegi Cymreictod.
- Llywodraeth Cymru a Senedd Cymru (polisi iaith): Mae CorCenCC yn hwyluso gwireddiad y pwyntiau gweithredu yn strategaeth Comisiynydd y Gymraeg sy'n ymwneud â chynnwys a chymwysiadau digidol, cyfieithu, terminoleg, cynllunio ieithyddol a gwaith ymchwil. Mae'r rhain yn adlewyrchu blaenoriaethau

Llywodraeth Cymru (2014; 2017). O ganlyniad, mae CorCenCC yn debygol o gael effaith ar ddatblygu integreiddiad y Gymraeg ym mywyd beunyddiol ymhellach, fel iaith a ddefnyddir ym meysydd llywodraethu a masnach, ac wrth ryngweithio'n gymdeithasol.

- Y diwydiant cyfieithu yng Nghymru: Mae allbynnau CorCenCC yn cyfateb â gwaith datblygu tymor canolig meddalwedd Microsoft Translate. Mae gwaith ymchwil rhagarweiniol (Screen, 2014) yn dangos y gall cyfieithu peirianyddol sy'n seiliedig ar enghreifftiau yn unig wella cynhyrchiant cyfieithwyr dynol hyd at 55%. Drwy gyfrannu at system cyfieithu peirianyddol hybrid yn y pen draw, gallai CorCenCC wella effeithlonrwydd cyfieithu ymhellach.
- Y cyfryngau yng Nghymru: Mae CorCenCC yn cynnig ffordd i gwmnïau'r cyfryngau werthuso natur ieithyddol eu hallbwn drwy fesur pa mor anodd yw eu deunydd i'w ddeall (e.e. cyfrifo cyfran yr eirfa sydd ag amlder isel) a thrwy ddarganfod Cymraeg pwy a gynrychiolir ac a dangynrychiolir. Drwy'r modd hwn, gallai CorCenCC gael effaith faterol ar hygyrchedd rhaglenni Cymraeg ac allbynnau'r cyfryngau, a pha mor ddeniadol ydynt, i'w cynulleidfaoedd targed, gan arwain at gynnydd mewn ffigurau gwylio/ymgysylltu o ganlyniad, yn ogystal â chefnogi cydraddoldeb cymdeithasol.
- Cyhoeddwyr a geiriadurwyr Cymraeg: Mae CorCenCC yn cynnig modd o dargedu cynnwys at gynulleidfaoedd sydd â gallu darllen gwahanol a gwella'r offerynnau ieithyddol sydd ar gael i awduron ar gyfer creu llyfrau darllen sydd wedi'u graddio. Bydd yn galluogi gwaith comisiynu geiriaduron Cymraeg modern sy'n seiliedig ar ddefnydd iaith gwirioneddol, gan gau'r bwlch a gydnabyddir, ac sy'n aml yn peri problemau, rhwng yr hyn sydd ei angen ar ddefnyddwyr y Gymraeg a'r hyn y gallant ddod o hyd iddo mewn ffynonellau cyfeiriadurol. Bydd hyn, yn ei dro, yn meithrin hyder siaradwyr Cymraeg yn eu galluoedd ieithyddol, gan eu gwneud yn fwy bodlon i ddefnyddio'r Gymraeg mewn amrediad ehangach o gyd-destunau.
- Cwmnïau technoleg ieithyddol: Mae corpws hyfforddi o'r radd flaenaf yn ofyniad allweddol ar gyfer cwmnïau sy'n defnyddio data'r cyfryngau cymdeithasol ar y we ac ar-lein, a dyma'r hyn y mae CorCenCC yn ei ddarparu. Bydd set ddata CorCenCC felly'n ei gwneud yn bosibl datblygu amrediad o adnoddau technoleg Cymraeg nad ydynt eto'n bodoli ar gyfer yr iaith.
- Y cyhoedd: Drwy ei gynllun a yrrir gan ddefnyddwyr, mae cynrychiolwyr y bobl sy'n debygol o ddefnyddio CorCenCC yn y dyfodol wedi bod ynghlwm wrth y gwaith o adeiladu a chynllunio'r corpws yn uniongyrchol, ac mae hyn wedi sicrhau ei fod yn hwylus i ddefnyddwyr, a'i fod yn hygyrch ac yn briodol i'w hanghenion. Bwriad y dull hwn yw adeiladu ar ddiddordeb sydd eisoes yn bodoli yn y Gymraeg a'i threftadaeth, a meithrin 'perchenogaeth' gymunedol o'r corpws. Mae'r effaith hirdymor bosibl yn cynnwys newid pendant mewn canfyddiadau o'r Gymraeg, ac ymagweddau tuag ati, yng Nghymru a'r tu hwnt.

9. Prosiectau cysylltiedig a chyllid pellach

Er y derbyniodd CorCenCC gyllid hael gan y Cyngor Ymchwil Economaidd a Chymdeithasol (ESRC) a'r Cyngor Ymchwil i'r Celfyddydau a'r Dyniaethau (AHRC),

cafodd amrediad o is-brosiectau a phrosiectau ymchwil cysylltiedig eraill eu hariannu gan ffynonellau eraill. Ceir manylion y rhain isod:

Dyddiad	Cyllidwr	Swm	Disgrifiad [ynghyd â'r prif ymchwilydd]
Ion 2017	Prifysgol Caerdydd	£56,000	Derbyniwyd cyllid gan Goleg y Celfyddydau, y Dyniaethau a'r Gwyddorau Cymdeithasol (AHSS) er mwyn cynnal ysgoloriaeth doethuriaeth tair blynedd ar gyfer Vigneshwaran Muralidaran ac astudiaeth yn dwyn y teitl '*Using insights from construction grammar for usage-based parsing*' [Knight a Spasić].
Chwef 2017	British Council	£2,000	Cyllid i gefnogi lansiad cyhoeddus prosiect CorCenCC yn Adeilad y Pierhead, Caerdydd [Knight].
Chwef 2017	Prifysgol Abertawe	£1,000	Derbyniwyd cyllid gan Sefydliad Ymchwil y Celfyddydau a'r Dyniaethau (RIAH) Prifysgol Abertawe i gefnogi lansiad prosiect CorCenCC [Fitzpatrick].
Chwef 2017	Prifysgol Caerdydd	£1,500	Derbyniwyd cymorth gan Gronfa Ymchwil ac Arloesedd Ysgolion ar gyfer lansiad prosiect CorCenCC [Knight].
Hyd 2017	Llywodraeth Cymru	£24,992	Comisiwn cystadleuol gan Lywodraeth Cymru i ddarparu asesiad cyflym o'r dystiolaeth o ymagweddau a dulliau effeithiol ar gyfer addysgu ail iaith. Am fwy o wybodaeth, gweler: https://tinyurl.com/ybtdsvfy [Fitzpatrick].
Ion 2018	Cynllun Grant Cymraeg 2050 2017-2018 GC2050/17-18/20:	£19,964	Cyllid ar gyfer adeiladu WordNet cyfrifiadurol ar gyfer y Gymraeg. Mae WordNet Cymru yn gronfa ddata eiriadurol lle caiff geiriau eu grwpio'n setiau o gyfystyron (synsetiau), sydd yno'n cael eu trefnu'n rhwydwaith o gysylltiadau semanteg-eiriadurol. I weld gwefan WordNet Cymru, ewch i: http://corcencc.org/wncy/ [Spasić].
Ion 2018	Cyd-bwyllgor Addysg Cymru (CBAC)	£1,968	Grant ymchwil (gan gynnwys rhaglen fewnfurol). Grant ymchwil i gwblhau gwaith ar lunio geirfa graidd B1 ar gyfer Cymraeg i Oedolion (lefel Canolradd). Am fwy o wybodaeth, ewch i: http://cronfa.swan.ac.uk/Record/cronfa48953 [Morris].
Maw 2018	Prifysgol Abertawe	£1,200	Lleoliad SPIN (interniaeth a delir gan Brifysgol Abertawe) ar gyfer gwaith casglu data, trawsgrifio a chyfweld ag athrawon/tiwtoriaid 2017-18. Ysgoloriaeth at ddibenion meithrin gallu [Morris].
Ebr 2018	Prifysgol Abertawe	£57,121	Cyllid gan Goleg y Celfyddydau a'r Dyniaethau (COAH) i gynnal ysgoloriaeth doethuriaeth tair blynedd ar gyfer Bethan Tovey-Walsh ac astudiaeth yn dwyn y teitl '*Purism and populism: The contested roles of code-switching and borrowing in minority language evolution*'. Talwyd ffioedd a chynhaliaeth [Morris a Fitzpatrick].
Gorff 2018	Prifysgol Caerdydd	£2,100	Cyllid mewnol CUROP (Cyfle Ymchwil Prifysgol Caerdydd) ar gyfer prosiect yn dwyn y teitl:

			'*Corpws Cenedlaethol Cymraeg Cyfoes: National Corpus of Contemporary Welsh – a focus on spoken data*'. Ysgoloriaeth at ddibenion meithrin gallu [Morris].
Gorff 2018	Prifysgol Caerdydd	£2,100	Cyllid mewnol CUROP (Cyfle Ymchwil Prifysgol Caerdydd) ar gyfer prosiect yn dwyn y teitl: '*Corpws Cenedlaethol Cymraeg Cyfoes: National Corpus of Contemporary Welsh – semantic tagging and data annotation*'. Ysgoloriaeth at ddibenion meithrin gallu [Morris].
Hyd 2018	Ysgoloriaeth gydweithredol ESRC DTP, Prifysgol Abertawe	£81,253	Y Gymraeg ac Ieithyddiaeth Gymhwysol Ysgoloriaeth Doethuriaeth Partneriaeth Hyfforddiant Doethurol (DTP) Cyngor Ymchwil Economaidd a Chymdeithasol (ESRC) Cymru, yn dwyn y teitl '*Strategic bilingualism: identifying optimal context for Welsh as a second language in the curriculum*' [Morris].
Ion 2019	Llywodraeth Cymru	£20,000	Cyllid i gefnogi'r gwaith o ddatblygu boniwr Cymraeg [Spasić].
Awst 2019	Llywodraeth Cymru	£90,000	Prosiect yn dwyn y teitl: '*Welsh language processing infrastructure: Welsh word embeddings*'. Mae ymgorffori geiriau'n math o gynrychioli geiriau lle caiff geiriau neu ymadroddion sydd ag ystyr tebyg eu mapio i fectorau o rifau real. Roedd y prosiect yn canolbwyntio ar ymgorffori geiriau ar gyfer y Gymraeg (ar greu geiriadur ac ymgorffori geiriau a thermau Cymraeg yn bennaf) ac mae'n cyfrannu at nod y Cynllun Gweithredu Technoleg Gymraeg i 'hybu adnoddau codio a thechnoleg Gymraeg ar gyfer athrawon a phlant ac eraill'.
Mai 2020	Llywodraeth Cymru	£90,000	Prosiect yn dwyn y teitl: 'Learning English-Welsh bilingual embeddings and applications in text categorisation'. Nod y prosiect hwn yw ymestyn canlyniadau'r prosiect ymgorffori geiriau blaenorol drwy greu cynrychioliadau trawsieithol o eiriau mewn man ymgorffori ar y cyd ar gyfer Cymraeg a Saesneg [Knight].
		£451,198	

10. Crynodeb o allbynnau'r prosiect

10.1. Offerynnau meddalwedd

Enw	Manylion	Dolen
Ap cyfrannu torfol CorCenCC:	Gyda'r nod o alluogi siaradwyr Cymraeg i recordio sgyrsiau rhyngddynt eu hunain a phobl eraill ar draws amrediad o gyd-destunau ac i'w lanlwytho, gan gynnwys caniatâd gan gyfranogwyr sy'n cydymffurfio'n foesegol, ar	http://www.corcencc.cymru/ap/ http://app.corcencc.org **Cyfeiriad:** Knight, D., Loizides, F., Neale, S., Anthony, L. a Spasić, I. (2020). Developing computational infrastructure for the CorCenCC corpus – the National Corpus of Contemporary

	gyfer eu cynnwys yn y corpws terfynol. Mae'r dull torfol o gasglu data corpws yn gyfeiriad datblygu cymharol newydd sy'n ategu dulliau mwy traddodiadol o gasglu data ieithyddol, ac mae'n cyfateb yn ddelfrydol â'r ysbryd cymunedol cadarnhaol sy'n bodoli ymhlith siaradwyr a defnyddwyr y Gymraeg.	Welsh. *Language Resources and Evaluation (LREV).*
CyTag – Tagiwr rhannau ymadrodd Cymraeg	Mae CyTag yn dagiwr Cymraeg arloesol (sy'n cynnwys set dagiau bwrpasol) a gafodd ei gynllunio a'i adeiladu ar gyfer y prosiect. Caiff ei ddefnyddio ar y cyd â'r tagiwr semantig i bennu tagiau i holl eitemau geiriadurol y corpws.	http://cytag.corcencc.org **Cyfeiriad:** 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. Poster a gyflwynwyd yn ystod *Cynhadledd Gwerthuso Adnoddau Iaith (LREC) 2018,* Mai 2018, Miyazaki, Japan.
Fersiwn 1 tagiwr semantig y Gymraeg CySemTag	Mae tagiwr semantig y Gymraeg yn cymhwyso anodiadau corpws i ddata am y Gymraeg mewn modd awtomataidd.	http://ucrel.lancs.ac.uk/usas/ **Cyfeiriad:** Piao, S., Rayson, P., Knight, D. a Watkins, G. 2018). Towards a Welsh Semantic Annotation System. *Proceedings of the LREC (Language Resources Evaluation 2018 Conference,* Mai 2018, Miyazaki, Japan. Piao, S., Rayson, P., Knight, D., Watkins, G. a Donnelly, K. (2017). Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language. *Proceedings of the Corpus Linguistics 2017 Conference,* Gorffennaf 2017, Prifysgol Birmingham, Birmingham, DU.
Seilwaith ac offerynnau ymholi CorCenCC	Mae offerynnau ymholi CorCenCC yn cynnwys y swyddogaethau canlynol: ▪ Ymholiad syml ▪ Ymholiad cymhleth ▪ Cynhyrchu rhestrau amlder ▪ Dadansoddi cydleoliad ▪ Dadansoddi n-gramau ▪ Cydgordio ▪ Dadansoddi geiriau allweddol	I weld yr offerynnau a chanllaw i ddefnyddwyr, ewch i: www.corcencc.cymru/archwilio **Cyfeiriad:** Knight, D., Loizides, F., Neale, S., Anthony, L. a Spasić, I. (2020). Developing computational infrastructure for the CorCenCC corpus – the National Corpus of Contemporary Welsh. *Language Resources and Evaluation (LREV).*
Y Tiwtiadur	Pecyn cymorth pedagogaidd CorCenCC sydd wedi'i integreiddio yn y prif offerynnau ymholi. Mae hyn yn cynnwys yr offerynnau dysgu ac addysgu canlynol: ▪ Llenwi bylchau ▪ Proffilydd geirfa	I weld yr offerynnau a chanllaw i ddefnyddwyr, ewch i: www.corcencc.cymru/archwilio **Cyfeiriad:** Davies, J., Thomas, E-M., Fitzpatrick, T., Needs, J., Anthony, L., Cobb, T. a Knight, D. (2020). *Y Tiwtiadur.* [Adnodd Digidol]. Ar gael yn: www.corcencc.cymru/Y-

	▪ Nodi geiriau ▪ Peiriant creu bylchau mewn brawddegau	tiwtiadur

10.2. Cyhoeddiadau (gyda'r dyddiadau yn y drefn wrthol, gydag enwau aelodau tîm y prosiect wedi'u nodi'n fras):

1. **Knight, D., Morris, S., Arman, L., Needs, J.** a **Rees, M.** (2021a, yn cael ei baratoi). *Blueprints for minoritised language corpus design: a focus on CorCenCC.* Llundain: Palgrave.
2. **Knight, D., Morris, S.** a **Fitzpatrick, T.** (2021b, yn cael ei baratoi). *Corpus Design and Construction in Minoritised Language Contexts:* **A focus on CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – National Corpus of Contemporary Welsh).** Llundain: Palgrave.
3. **Knight, D.,** Loizides, F., **Neale, S., Anthony, L.** a **Spasić, I.** (2020). Developing computational infrastructure for the CorCenCC corpus – the National Corpus of Contemporary Welsh. **Language Resources and Evaluation (LREV).**
4. Corcoran, P., Palmer, G., **Arman, L., Knight, D.** a **Spasić, I.** (2020, derbyniwyd). Word Embeddings in Welsh. *Journal of Information Science*.
5. **Muralidaran, V., Knight, D.** a **Spasić, I**. (2020, derbyniwyd). A systematic review of unsupervised approaches to usage-based grammar induction. *Natural Language Engineering.*
6. **Spasić, I.**, Owen, D., **Knight, D.** ac Arteniou, A. (2019). Data-driven terminology alignment in parallel corpora. *Proceedings of the* **Celtic Language Technology Workshop 2019,** Dulyn, Iwerddon.
7. **Piao, S., Rayson, P., Knight, D.** a **Watkins, G**. 2018). Towards a Welsh Semantic Annotation System. *Proceedings of the LREC (Language Resources Evaluation) 2018 Conference,* Mai 2018, Miyazaki, Japan.
8. **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. Poster a gyflwynwyd yn ystod *Cynhadledd Gwerthuso Adnoddau Iaith (LREC) 2018,* Mai 2018, Miyazaki, Japan.
9. **Rayson, P.** 2018). Increasing Interoperability for Embedding Corpus Annotation Pipelines in Wmatrix and other corpus retrieval tools. Trafodion y gweithdy Heriau wrth Reoli Corpysau Mawr yn ystod *Cynhadledd Gwerthuso Adnoddau Iaith (LREC) 2018,* Mai 2018, Miyazaki, Japan.
10. **Rayson, P**. a **Piao, S**. (2017). Creating and Validating Multilingual Semantic Representations for Six Languages: Expert versus Non-Expert Crowds. Trafodion y Gweithdy Cyntaf ar 'Sense, Concept and Entity Representations and their Applications' a gynhaliwyd yn ystod cynhadledd yr *European Chapter of the Association for Computational Linguistics 2017* (EACL), Ebrill, Valencia.
11. **Piao, S., Rayson, P.,** Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R-M., **Knight, D.,** Křen, M., Löfberg, L., Nawab, R. M. A., Shafi, J., Teh, P-L., a Mudraya, O. (2016). Lexical Coverage Evaluation of Large-scale Multilingual

Semantic Lexicons for Twelve Languages. *Proceedings of the LREC (Language Resources Evaluation) 2016 Conference,* Mai 2016, Miyazaki, Slofenia.

10.3. Prif gyflwyniadau a gwahoddiadau i siarad

Mae gwaith ymchwil prosiect CorCenCC wedi'i gyflwyno yn ystod 17 o brif gyflwyniadau a gwahoddiadau i siarad, ac mae wedi'i ledaenu drwy 37 o bapurau cynhadledd eraill mewn 11 o wledydd o gwmpas y byd. Gellir cael manylion am y digwyddiadau siarad hyn ar brif wefan CorCenCC (gweler: www.corcencc.cymru/allbynnau).

Cyfeiriadau

Aston, G. (2001) *Learning with Corpora*, Athelstan, Open Library.

Aston, G. a Burnard, L. (1997) *The BNC Handbook: Exploring the British National Corpus with SARA,* Caeredin: Gwasg Prifysgol Caeredin.

Ball, M. a Müller, N. (1992) *Mutation in Welsh,* Clevedon: Multilingual Matters.

Brabham, D. C. (2008) 'Crowdsourcing as a model for problem solving: An introduction and cases', *Convergence* 14**:** 75-90.

Carter, R. a McCarthy, M. (2004) 'Talking, creating: Interactional language, creativity, and context', *Applied Linguistics* 25**:** 62-88.

Cobb, T. (2000). *The compleat lexical tutor* [Ar-lein], ar gael: http://www.lextutor.ca/ [Cyrchwyd 07/07/20].

Collins. (2020). *Collins Corpus online* [Ar-lein], ar gael: https://collins.co.uk/pages/elt-cobuild-reference-the-collins-corpus [Cyrchwyd 07/07/20].

Cooper, Jones, D. a Prys (2019) 'Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology', *Information* 10**:** 247.

Cambridge University Press. (2020) *Cambridge English Corpus online* [Ar-lein], ar gael: https://www.cambridge.org/us/cambridgeenglish/better-learning-insights/corpus [Cyrchwyd 07/07/20].

Deuchar, D., Webb-Davies, P. a Donnelly, K. (2018) *Building and Using the Siarad Corpus,* Amsterdam: John Benjamins.

Deuchar, M., Davies, P., Herring, J. R., Parafita Couto, M. a Carter, D. (2014) 'Building bilingual corpora: Welsh-English, Spanish-English and Spanish-Welsh', yn Thomas, E. M. a Mennen, I. (gol.) *Advances in the Study of Bilingualism.* Bristol: Multilingual Matters.

Donnelly, K. (2013a) *Eurfa v3.0 - Free (GPL) Dictionary (incorporating Konjugator and Rhymer)* [Ar-lein], ar gael: http://eurfa.org.uk [Cyrchwyd 07/07/20].

Donnelly, K. (2013b) *Kynulliad3: a corpus of 350,000 aligned Welsh and English sentences from the Third Assembly (2007-2011) of the National Assembly for Wales* [Ar-lein], ar gael: http://cymraeg.org.uk/kynulliad3/ [Cyrchwyd 07/07/20].

Donnelly, K. a Deuchar, M. (2011) 'The Bangor Autoglosser: A multilingual tagger for conversational text', yn Proceedings of *the Fourth International Conference on Internet Technologies and Applications (ITA11),* Wrecsam, Cymru. tt. 17-25.

Expert Advisory Group on Language Engineering Standards. (1996) *EAGLES guidelines* [Ar-lein], ar gael: http://www.ilc.cnr.it/EAGLES/browse.html [Cyrchwyd 07/07/20].

Estellés-Arolas, E. a González-Ladrón-De-Guevara, F. (2012) 'Towards an integrated crowdsourcing definition', *Journal of Information Science* 38**:** 189-200.

Evas, J. a Williams, C. H. (1998) 'Community language regeneration: realising potential', yn Nhrafodion yr *International Conference on Community Language Planning,* Caerdydd, Bwrdd yr Iaith Gymraeg. tt. 1-13.

Hardie, A. (2012) 'CQPweb – combining power, flexibility and usability in a corpus analysis tool', *International Journal of Corpus Linguistics* 17**:** 380-409.

Hawtin, A. (2018) *The Written British National Corpus 2014: Design, compilation and analysis,* Traethawd PhD heb ei gyhoeddi: Prifysgol Caerhirfryn.

Johns, T. (1991) 'Should you be persuaded: Two samples of data-driven learning materials', *English Language Research Journal* 4: 1-16.

Karlsson, F. (1990) 'Constraint grammar as a framework for parsing running text', yn Nhrafodion y *13th International Conference on Computational Linguistics (COLING),* Helsinki, Y Ffindir. tt. 168-173.

Karlsson, F., Voutilainen, A., Heikkilä, J. a Anttila, A. (1995) *Constraint grammar: A language-independent framework for parsing unrestricted text,* Berlin/Efrog Newydd: Mouton de Gruyter.

Knight, D., Adolphs, S. a Carter, R. (2013) 'Formality in digital discourse: a study of hedging in CANELC', in Romero-Trillo, J. (gol.) *Yearbook of corpus linguistics and pragmatics,* Yr Iseldiroedd: Springer. tt. 131-152.

Leńko-Szymańska, A. a Boulton, A. (2015) *Multiple Affordances of Language Corpora in Data-driven Learning,* Amsterdam: John Benjamins.

Little, D. (2007) 'Language learner autonomy: Some fundamental considerations revisited', *Innovations in Language Learning and Teaching* 1: 14-29.

Llywodraeth Cymru (2013) *Codi golygon: adolygiad o Gymraeg i Oedolion. Adroddiad ac argymhellion [Raising our sights: review of Welsh for Adults. Report and recommendations]*, Bedwas: Llywodraeth Cymru.

Llywodraeth Cymru (2014) *Iaith fyw: iaith byw - Bwrw mlaen*, Caerdydd: Llywodraeth Cymru.

Llywodraeth Cymru (2017) *Cymraeg 2050: Miliwn o siaradwyr Cynllun Gweithredu 2019-20*, Caerdydd: Llywodraeth Cymru.

McEnery, T., Love, R., & Brezina, V. (2017) 'Compiling and analysing the Spoken British National Corpus 2014', *International Journal of Corpus Linguistics 22*(3): 311-318.

Mac Giolla Chríost, D., Carlin, P., Davies, S., Fitzpatrick, T., Jones, A. P., Heath-Davies, R., Marshall, J., Morris, S., Price, A., Vanderplank, R., Walter, C. a Wray, A. (2012). *Adnoddau, dulliau ac ymagweddau dysgu ac addysgu ym maes Cymraeg i Oedolion: astudiaeth ymchwil gynhwysfawr ac adolygiad beirniadol o'r ffordd ymlaen [Welsh for Adults teaching and learning approaches, methodologies and resources: a comprehensive research study and critical review of the way forward]*, Bedwas: Llywodraeth Cymru.

Nation, I.S.P. (2001) *Learning Vocabulary in Another Language*, Caergrawnt: Gwasg Prifysgol Caergrawnt.

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' yn Nhrafodion yr *Eleventh International Conference on Language Resources and Evaluation (LREC),* Miyazaki, Siapan. tt. 3946-3954.

NFER (2008) *Ymchwil i'r Cwrs Dwys ar gyfer Cymraeg i Oedolion*, Abertawe: Sefydliad Cenedlaethol er Ymchwil i Addysg.

ONS (2011) *DC2612WA – Ability to speak Welsh by occupation* [Ar-lein], Office for National Statistics, Durham: Nomis. Ar gael: www.nomisweb.co.uk/census/2011/dc2612wa [Cyrchwyd 07/07/20].

Piao, S., Rayson, P., Knight, D. a Watkins, G. (2018) 'Towards a Welsh semantic annotation system' yn Nhrafodion yr *Eleventh International Conference on Language Resources and Evaluation (LREC),* Miyazaki, Siapan. tt. 980-985.

Rayson, P., Archer, D., Piao, S. a McEnery, T. (2004) 'The UCREL semantic analysis system', yn Nhrafodion y *Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP tasks at the 4th International Conference on Language Resources and Evaluation (LREC),* Lisbon, Portiwgal. tt. 1-6.

Scannell, K. (2007) 'The Crúbadán Project: Corpus building for under-resourced languages' yn *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*. tt. 1-10.

Scannell, K. (2012) *Kevin Scannell's website* [Ar-lein], ar gael: http://borel.slu.edu/ [Cyrchwyd 07/07/20].

Sinclair, J. (2005) 'Corpus and text - basic principles', yn Wynne, M. (gol.) *Developing Linguistic Corpora: a Guide to Good Practice,* Rhydychen: Oxbow Books. tt. 1-16.

Thomas, E. M. a Mayr, R. (2010) 'Children's acquisition of Welsh in a bilingual setting: a psycholinguistic perspective, yn Morris, D. (gol.) *Welsh in the 21st Century,* Caerdydd: Gwasg Prifysgol Caerdydd.

Thompson, P. (2005) 'Spoken Language Corpora' yn Wynne, M. (gol.) *Developing Linguistic Corpora: a Guide to Good Practice,* Rhydychen: Oxbow Books. tt. 59-70.

Wilson, A. (2002) *The Language Engineering Resources for the Indigenous Minority Languages of the British Isles and Ireland Project* [Ar-lein], ar gael: https://www.lancaster.ac.uk/fass/projects/biml/default.htm [Cyrchwyd 07/07/20].