

Iaith a Thechnoleg yng Nghymru: Cyfrol I

Golygydd: Delyth Prys



PRIFYSGOL
BANGOR
UNIVERSITY



Cyhoeddwyd yr e-lyfr hwn gyntaf yn 2021 gan

Prifysgol Bangor, Ffordd y Coleg, Bangor, Gwynedd LL57 2DG

www.bangor.ac.uk/cy

Rhif Llyfr Rhyngwladol (e-lyfr):

ISBN: 978-1-84220-189-3

Mae'r testun wedi'i ryddhau dan y drwydded Creative Commons BY 4.0

<https://creativecommons.org/licenses/by/4.0/>, caniatáu i chi ei aildddefnyddio a'i newid mewn unrhyw ffordd os ydych yn rhoi cydnabyddiaeth briodol. Gweler testun y drwydded <https://creativecommons.org/licenses/by/4.0/> am ragor o fanylion.

Cymorth dylunio a phrawfddarllen gan Stefano Ghazzali a Dr Gareth Watkins.

Mae'r llyfr hwn ar gael hefyd yn Saesneg dan y teitl *Language and Technology in Wales Volume I*, rhif ISBN 978-1-84220-188-6.

Iaith a Thechnoleg yng Nghymru: Cyfrol 1

Golygydd

Delyth Prys PRIFYSGOL BANGOR

Cyfranmyr

Kepa Sarasola PRIFYSGOL GWLAD Y BASG

Iñaki Alegria PRIFYSGOL GWLAD Y BASG

Olatz Perez-de-Viñaspre PRIFYSGOL GWLAD Y BASG

Geraint Palmer PRIFYSGOL CAERDYDD

Padraig Corcoran PRIFYSGOL CAERDYDD

Laura Arman PRIFYSGOL CAERDYDD

Dawn Knight PRIFYSGOL CAERDYDD

Irena Spasic PRIFYSGOL CAERDYDD

Dewi Bryn Jones PRIFYSGOL BANGOR

Sarah Cooper PRIFYSGOL BANGOR

Myfyr Prys CYMEN CYF

Vigneshwaran Muralidaran PRIFYSGOL CAERDYDD

Keeziah O'Hare PRIFYSGOL CAERDYDD

Gruffudd Prys PRIFYSGOL BANGOR

Gareth Watkins PRIFYSGOL BANGOR

Jonathan C Roberts PRIFYSGOL BANGOR

Peter W. S Butcher PRIFYSGOL BANGOR

Robert Lew PRIFYSGOL ADAM MICKIEWICZ

Geraint Rees PRIFYSGOL SURREY,

Nirwan Sharma PRIFYSGOL AGORED

Ana Frankenberg-Garcia PRIFYSGOL SURREY,

Leena Sarah Farhat PRIFYSGOL BANGOR

William John Teahan PRIFYSGOL BANGOR

Cynnwys

Rhagair	
JEREMY MILES AS, GWEINIDOG Y GYMRAEG AC ADDYSG	5
Rhagymadrodd	
DELYTH PRYS, PRIFYSGOL BANGOR	6
1 Technoleg Iaith ar gyfer Cymunedau Ieithyddol: Trosolwg yn Seiliedig ar Ein Profiadau	
KEPA SARASOLA, IÑAKI ALEGRIA, OLATZ PEREZ-DE-VIÑASPRE	9
2 Golwg Agosach ar Fewnblaniadau Geiriau Cymraeg	
GERAINT PALMER, PADRAIG CORCORAN, LAURA ARMAN, DAWN KNIGHT, IRENA SPASIC	20
3 Gweithredu Boniwr Porter ar gyfer y Gymraeg	
VIGNESHWARAN MURALIDARAN, GERAINT PALMER, LAURA ARMAN, KEEZIAH O'HARE, DAWN KNIGHT, IRENA SPASIC	29
4 Datblygu ColloCaid: offer delweddu a golygu testun i helpu ysgrifenyddwyr â chydosiadau	
JONATHAN C ROBERTS, PETER W. S BUTCHER, ROBERT LEW, GERAINT REES, NIRWAN SHARMA, ANA FRANKENBERG-GARCIA	42
5 Modelu a Phrosesu Testun Cymraeg gan ddefnyddio Pecyn Cymorth TAWA	
WILLIAM JOHN TEAHAN, LEENA SARAH FARHAT	55
6 Testun-i-Leferydd Ddwyieithog Cymraeg a Saesneg	
DEWI BRYN JONES, SARAH COOPER	63
7 Datblygu Tagiwr rhannau ymadrodd a Chorpws o Frawddegau Hyfforddi ar gyfer y Gymraeg	
GRUFFUDD PRYS, GARETH WATKINS	74
8 Model Word2vec Cymraeg: cynrychioliad fector o gydberthynas semantig geiriau Cymraeg	
ar sail eu mewnbaniadau o fewn corpwys enfawr Cymraeg	
GRUFFUDD PRYS, GARETH WATKINS	86
9 Gwreiddio NMT mewn cwmni cyfieithu Cymraeg	
MYFYR PRYS	107

Rhagair

JEREMY MILES AS, GWEINIDOG Y GYMRAEG AC ADDYSG

Pan gefais fy mhenodi'n Weinidog y Gymraeg ac Addysg, dywedais fy mod am ganolbwyntio ar gynyddu'r defnydd o'n hiaith. Ac mae rôl ganolog gan dechnoleg yn hynny o beth – yn arbennig yng ngoleuni COVID-19 sydd wedi newid bywydau cynifer ohonom dros y cyfnod diweddar.

Wrth edrych *Gynllun Gweithredu Technoleg Gymraeg* Llywodraeth Cymru, fe welwch fod ein ffocws ar dechnolegau lleferydd, cyfieithu a deallusrwydd artiffisial. Rydym wedi trefnu fod cymaint o'r gwaith ag sy'n bosibl ar gael i'w ddefnyddio a'i ailddefnyddio yn rhad ac am ddim, o dan drwydded agored. Mawr obeithiaf y bydd hyn yn ei wneud yn ddeniadol ac yn hawdd i gwmnïau eu mabwysiadu yn eu cynnyrch—a chynnig y cynnyrch hwnnw yn rhagweithiol yn Gymraeg i gwsmeriaid, cleientiaid a dinasyddion. Gwneud y Gymraeg yn haws: dyna'r tric wrth gynyddu defnydd. A dyna nod y gwaith rydym yn ei wneud yn Llywodraeth Cymru er mwyn ymgorffori Profiad Dwyieithog Da yn ein systemau caffael technoleg.

Mae'r ymchwil a'r gwaith manwl y byddwch yn darllen amdanynt yn y gyfrol hon wedi creu adnoddau y gallwn ni gyd ymfalchïo ynddynt. A gall ein cydweithwyr mewn cymunedau ieithyddol eraill hefyd fanteisio ar ffrwyth y llafur.

Mae technoleg yn cyson esblygu, a rhaid i'r Gymraeg hithau esblygu gyda'r dechnoleg. Edrychaf ymlaen at weld rhagor o ddatblygiadau yn y maes.

Diolch i'r holl ddatblygwyr a phawb fu ynghlwm wrth y gwaith.

Rhagymadrodd

DELYTH PRYS - PRIFYSGOL BANGOR

Mae'r gyfrol hon yn seiliedig ar rai o'r prif bapurau a draddodwyd yn ystod Symposiwm Academiaidd Technolegau Iaith Cymru 2020. Cynhaliwyd y symposiwm ar gais a chyda chefnogaeth Llywodraeth Cymru, a'r bwriad oedd dod â rhai o'r prif ymchwilwyr y maes yng Nghymru at ei gilydd, rhannu gwybodaeth a hybu ymchwil pellach. Oherwydd y cyfyngiadau Covid, bu'n rhaid cynnal y Symposiwm ar ffurf rithiol, ac mae hynny ynddo'i hun yn dweud llawer am bwysigrwydd technoleg yn y byd cyfoes, a'i allu i hwyluso cyfathrebu rhyngom â'n gilydd.

Mae Technolegau Iaith yn cynnwys technolegau lleferydd (pethau fel trawsgrifio darlith neu sgwrs o'r llafar i'r ysgrifenedig, a chreu lleisiau synthetig), technolegau testun (sy'n cynnwys gwirwyr gramadeg ac adnabod emosiwn neu ystyr mewn testunau), prosesu iaith naturiol (er enghraifft dadansoddi cystrawen a morffoleg, alldynnu termau ac anonymeiddio testun). Erbyn hyn defnyddir dulliau niwral Dysgu Dwfn yn y technolegau hyn i gyd, ac mae'n cael llawer gwell canlyniadau na'r dulliau oedd mewn bri cyn hynny. Ar y llaw arall, mae hyn peri problem ychwanegol i ieithoedd llai eu hadnoddau fel y Gymraeg. Er mwyn gallu defnyddio technegau niwral newydd, mae angen meintiau anferthol o ddata o wahanol fathau, ar ffurf testun neu leferydd, er mwyn canfod y patrymau angenrheidiol a hyfforddi modelau iaith. Dyma'r union ddata sydd yn brin mewn ieithoedd llai eu hadnoddau. Mae'r ieithoedd hyn yn cynnwys rhai rydym wedi arfer â meddwl amdanynt nhw fel ieithoedd 'lleiafrifol', ieithoedd heb lawer o siaradwyr, a ieithoedd cymunedau difreintiedig, fel rhai o ieithoedd India ac Affrica. Gall ieithoedd y cymunedau difreintiedig hyn fod â miliynau o siaradwyr, ond efallai bod llythrennedd yn isel yn eu plith, neu nad oes cyflenwad trydan a rhynggrwyd dibynadwy ganddynt – rhai o hanfodion cyfathrebu cyfoes.

Ystyrir bellach bod cael offer cyfathrebu ar sail technolegau iaith yn allweddol ar gyfer parhad a ffyniant ieithoedd dynol. Heb raglenni cyfrifiadurol sy'n medru delio gyda'ch iaith, does dim modd defnyddio'r amrediad cyflawn eich iaith mewn cyd-destunau digidol ac ar y we fyd-eang, a rhaid troi at un o'r ieithoedd mawr i roi gorchmynion llafar i'ch teclynau electronig, darllen testunau, neu gael cymorth i ysgrifennu'n gywir. Mae perygl difodiant i ieithoedd llai eu hadnoddau felly yn un real, ac mae pwnc adfywio ieithoedd drwy dechnolegau iaith wedi dod yn un pwysig i gynllunwyr ieithyddol a pholisi llywodraethau sy'n poeni am ddiogelu ieithoedd a hawliau sifil eu cymunedau hwy. Dyna paham mae ymchwilwyr yng Nghymru yn ceisio mynd i'r afael â'r heriau hyn, a chynnig ateb ar gyfer y Gymraeg, ac ieithoedd eraill mewn sefyllfaoedd tebyg.

Un cwestiwn ymchwil diddorol yn y maes yw a oes modd datblygu technolegau iaith ar gyfer yr ieithoedd hyn yn yr un ffordd ag y gwneir ar gyfer ieithoedd mawr y byd? Rhan bwysig o hyn yw canfod sut mae cael hyd i ddata digonol mewn ieithoedd llai eu hadnoddau er mwyn bodloni gofynion Deallusrwydd Artiffisial a Dysgu Dwfn. Mae gwirfoddolwyr yng Nghymru wedi gwneud gwyrdoliau mewn projectau fel Common Voice gan Mozilla yn recordio eu lleisiau er mwyn cael digon o ddata lleferydd, ac mae parodrwydd cyrff cyhoeddus i rannu testunau megis eu cofion cyfieithu hefyd yn rhan o'r ateb. Atebion posibl eraill yw defnyddio dulliau cyfieithu peiranyddol i greu 'data ffug' lle nad oes llawer o ddata go iawn ar gael mewn iaith, yn enwedig ar gyfer rhai pynciau, a defnyddio dulliau dysgu trosglwyddol rhwng iaith lle datblygwyd rhyw dechnoleg eisoes a iaith arall lle na cheir y dechnoleg honno eto. Rydym yng nghanol y berw syniadol, arloesol hwn ar hyn o bryd, ac mae'n amser cyffrous iawn i fod yn ymchwilydd yn y maes.

Yng Nghymru mae ymchwilwyr wedi bod yn weithgar yn y maes ym Mhrifysgol Bangor ers nifer o flynyddoedd, a chafwyd gweithgaredd mwy diweddar hefyd ym Mhrifysgolion Caerdydd a De Cymru. Mae gennym gymuned weithgar o wirfoddolwyr, sy'n creu a chyfrannu data Cymraeg, a hefyd sylfaen o gwmnïau bach a chanolig lleol, fel cwmni cyfieithu Cymen yng Nghaernarfon, sy'n ymgorffori ffrwyth yr ymchwil hwn yn eu cynnyrch, a thrwy hynny yn hybu economi a llesiant cymdeithasol Cymru. Rydym wedi dechrau arni i wella addysg technolegau iaith yng Nghymru, ac mae'r cwrs MSc newydd mewn Technolegau Iaith ym Mhrifysgol Bangor yn gam i'r cyfeiriad cywir, ond mae llawer eto i'w wneud, gan ddechrau gyda chynyddu'r pwyslais ar ddysgu codio i blant yn ein hysgolion cynradd ac uwchradd, a dod â'r elfennau ieithyddol a chyfrifiadurol at ei gilydd ar lefel israddedig ac ôl-raddedig yn ein prifysgolion.

Mae ymchwil technolegau iaith yn ei hanfod yn weithgaredd trawsbynciol, ac er mwyn cryfhau'r cydweithio rhwng gwahanol ymchwilwyr a rhanddeiliaid eraill yng Nghymru, fe sefydlwyd Rhwydwaith Technolegau Iaith Cenedlaethol yn 2020 (<https://rhwydwaith.techiaith.cymru/>). Mae croeso i unrhyw un ymuno â'r rhwydwaith hon, ac mae'n fan cyfarfod a chyfnewid syniadau nid yn unig i academyddion o wahanol adrannau a phrifysgolion, ond hefyd i bobl o fyd diwydiant a byd polisi, heb anghofio'r llu gwirfoddolwyr sydd mor bwysig i alluogi iaith fach fel y Gymraeg i ddatblygu adnoddau ac offer technolegau iaith blaengar iddi ei hun. Mae creu technolegau iaith hefyd yn weithgaredd byd-eang, o ddiddordeb i gorfforaethau mawr rhyngwladol fel Google a Microsoft, yn ogystal â gwladwriaethau, cymunedau diwladwriaeth ac ymgyrchwyr iaith. Yn sgil cynnig llwyddiannus Jill Evans ASE yn Senedd Ewrop ym Medi 2018 ar gydraddoldeb ieithyddol yn yr oes ddigidol (https://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html), aeth y Comisiwn Ewropeaidd ati i greu project i lunio agenda a map ffordd er mwyn ennill cydraddoldeb digidol llawn i holl ieithoedd Ewrop erbyn 2030 (<https://european-language-equality.eu/>). Golyga hyn bwyslais newydd ar ddatblygu technolegau iaith addas yn ystod y blynyddoedd nesaf, ac mae angen i ymchwilwyr Cymru mewn prifysgolion a diwydiant fod yn barod ar gyfer hynny.

Mae'r gyfrol hon yn gyfraniad tuag at ddatblygu'r maes yng Nghymru. Fe'i cyhoeddir yn ddwyieithog dan drwydded agored ganiataol a fydd yn golygu bod testun cyfochrog y gyfrol yn Gymraeg ac yn Saesneg yn cael ei ychwanegu at gorpws trwydded ganiataol Bangor ac yn cyfrannu at y Data Mawr yma sydd mor bwysig i ddatblygiad technolegau iaith. Mae'r gyfrol yn cynnwys yn gyntaf bapur yn adrodd ar y sefyllfa wrth geisio datblygu technolegau iaith ar gyfer y Fasgeg, iaith fach arall mewn sefyllfa debyg iawn i'r Gymraeg, gan Kepa Sarasola, Iñaki Alegria, Olatz Perez-de-Viñaspre o Brifysgol Gwlad y Basg. Mae'r papur hwn yn rhoi cyd-destun ehangach i'n hymdrechion ni yng Nghymru, ac yn ein hatgoffa o werth cydweithio a rhannu syniadau ar draws ffiniau daearyddol. Mae papurau eraill yn y gyfrol gan ymchwilwyr o sawl adran ym Mhrifysgol Caerdydd sydd wedi bod yn ymchwilio i fewnblaniadau geiriau Cymraeg a gweithredu boniwr Saesneg Porter ar gyfer y Gymraeg, sef Geraint Palmer, Padraig Corcoran, Laura Arman, Dawn Knight, Irena Spasic a Vigneshwaran Muralidaran, Laura Arman, a Keeziah O'Hare. Edrych ar addasu rhaglenni a ddatblygwyd yn wreiddiol ar gyfer y Saesneg, ac yn benodol y rhaglen Collocaid mae Jonathan C Roberts a Peter W. S. Butcher o Brifysgol Bangor, ar y cyd gyda Robert Lew o Brifysgol Adam Mickiewicz, Nirwan Sharma o'r Brifysgol Agored a Geraint Rees o Brifysgol Surrey, y DU, ac Ana Frankenberg-Garcia, ill dau o Brifysgol Surrey. Mae Bill Teahan a Leena Farhat o Brifysgol Bangor yn dilyn trywydd tebyg wrth drafod addasu pecyn cymorth TAWA ar gyfer y Gymraeg. O Brifysgol Bangor y daw hefyd papurau gan Dewi Jones a Sarah Cooper yn trafod testun-i-leferydd dwyieithog, a Gruffudd Prys a Gareth Watkins yn trafod llyfrgell newydd ar gyfer prosesu'r Gymraeg ar-lein. Mae papur Myfyr Prys,

cyn ymchwilydd KTP a Smart rhwng Prifysgol Bangor a chwmni Cymen, ond sydd bellach yn gweithio i Cymen, yn trafod datblygu cyfieithu peirianyddol Saesneg<>Cymraeg ar gyfer cwmni cyfieithu.

Dyma gipolwg felly ar rai o'r pynciau ymchwil cyfoes sy'n mynd â bryd technolegwyr iaith yng Nghymru ac Ewrop ar hyn o bryd. Y gobaith yw y bydd y papurau hyn yn fuddiol i hybu rhagor o ymchwil yn y maes, i addysgu myfyrwyr, ac i gyfrannu at ddatblygu offer technoleg iaith er budd y Gymraeg a icithoedd eraill llai eu hadnoddau.

Technoleg iaith ar gyfer Cymunedau ieithyddol: Trosolwg yn Seiliedig ar Brofiadau'r Fasgeg

KEPA SARASOLA

Grŵp IXA, Prifysgol Gwlad y Basg (UPV/EHU)

IÑAKI ALEGRIA

Grŵp IXA, Prifysgol Gwlad y Basg (UPV/EHU)

OLATZ PEREZ-DE-VIÑASPRE

Grŵp IXA, Prifysgol Gwlad y Basg (UPV/EHU)

Mae IXA yn grŵp ymchwil sydd wedi bod yn gweithio ar dechnolegau iaith, yn y Fasgeg yn bennaf, dros y 32 blynedd diwethaf. Yn sgil blyneddau o gydweithio gyda'r gymuned Fasgeg a chymunedau ieithyddol eraill, rydym wedi dod i weld technoleg fel arf hanfodol. Ond mae angen rhywfaint o waith craidd cychwynnol cyn defnyddio'r technegau hyn: 1) safoni a 2) cynhyrchu cynnwys agored. Gan gadw'r hanfodion hyn mewn cof, rydym yn awgrymu sefydlu pecynnau BLARK (Basic Language Resource Kit) i adnabod set o adnoddau craidd, ac yn awgrymu offer ar gyfer addasu'r rhain i wahanol ieithoedd gan ddibynnu ar hyd a lled yr adnoddau cymunedol a digidol sydd ar gael i'w siaradwyr. I orffen, byddwn yn cyflwyno detholiad o arbrofion llwyddiannus sy'n ein galluogi i fod yn obeithiol am y defnydd posibl o rwydweithiau niwral, technegau dysgu dwfn a modelau BERT gydag ieithoedd llai eu hadnoddau.

1 CYFLWYNIAD

Cafodd y grŵp ymchwil IXA (www.ixa.eus) ei greu yn 1988 gyda'r nod o osod sylfeini ymchwil a datblygu Prosesu Testunau Iaith Naturiol (NLP) a Thechnoleg Iaith Ddynol (HLT) ar gyfer y Fasgeg. Mae bellach yn grŵp amlddisgyblaethol sylweddol sy'n cynnwys gwyddonwyr cyfrifiadurol ac ieithyddion.

Dwy nodwedd hynod sy'n perthyn i IXA yw ei fod yn ymdrin ag iaith lai ei hadnoddau (y Fasgeg) a'i fod yn cyfuno modelu ieithyddol clasurol, dadansoddi data gyda dulliau NLP tebygolïaethol a dysgu peirianyddol arloesol.

Ar y dechrau un, tri deg naw o flynyddoedd yn ôl, arweiniodd ein grant cyntaf at greu system cyfieithu ar gyfer y pâr ieithyddol Sbaeneg-Basgeg. Ond ar ôl rhai astudiaethau rhagarweiniol, sylweddolwyd ei bod yn bwysicach canolbwyntio ein hymdrechion ar greu adnoddau ac offer sylfaenol ar gyfer y Fasgeg (offeryn dadansoddi a chynhyrchu morffoleg, geiriaduron electronig, corpora anodedig, cronfeydd data semantig, ac ati) y gellid eu

defnyddio'n ddiweddarach i greu llawer o raglenni ieithyddol cyffredinol eraill, yn hytrach na chreu system cyfieithu peirianyddol a fyddai'n un *ad hoc* ac eithriadol o gymhleth. Dyma'r ystyriaeth a oedd yn greiddiol i'n strategaeth i wneud cynnydd o ran datblygu technoleg iaith i'r Fasgeg.

Erbyn hyn, mae ein hymchwyl wedi arwain at ddatblygu technolegau prosesu iaith naturiol (NLP) o'r radd flaenaf ar gyfer y Fasgeg, technoleg sy'n gadarn ac yn eang ei chwmpas. Ymysg y technolegau a'r adnoddau hyn mae'r gwirydd sillafu (Xuxen), rhwyd geiriau Basgeg (BasqueWN), corpws gwyddoniaeth a thechnoleg (corpws ZT), corpws wedi'i anodi'n gystrawennol (EPEC), system cyfieithu peirianyddol Spaeneg-Basgeg (Matxin), Piblinell NLP ar gyfer prosesu testun (IXA-pipes) ac offeryn cloddio safbwyntiau (Behagunea).

Yn seiliedig ar ein profiad gydag NLP ar gyfer ieithoedd llai eu hadnoddau [1], rydym wedi bod yn cydweithio ers blyneddau llawer â dau fath o gymuned ieithyddol:

- Y rheini sy'n gweithio ar gymdeithasoli'r Fasgeg (gan ddatblygu geiriaduron, dulliau dysgu iaith, Wikipedia, bysellfyrddau ac offer dehongli ar gyfer ffonau clyfar...)
- Y rheini sy'n gweithio ar ieithoedd prin eu hadnoddau, er mwyn eu helpu i ddatblygu eu hiaith yn dechnolegol (Cetshwa, Nahuatl, Sbaeneg Ciwba...).

Cyfeiriodd Borin [2] at addewid HLT i ieithoedd llai adnabyddus, ac mae'n disgrifio'r amrywiaeth ieithyddol yn y gymdeithas wybodaeth. Mae'n dyfynnu Ostler: *"a language will not get on in the world of today unless it is supplied with a distributor and a multi-word corps of text"*. Dadansoddodd Ostler y berthynas rhwng cymdeithaseg iaith ac HLT, a chynhigiodd rai ystyriaethau strategol rydym wedi'u mabwysiadu.

Yn ein barn ni, gall technoleg fod yn ffactor pwysig o ran datblygu iaith, ond mae rhywfaint o waith craidd y mae'n rhaid ei weithredu ymlaen llaw (neu ar yr un pryd):

- Safoni: mae darnio'r gymuned yn dafodieithoedd yn gwneud cynhyrchu testunau ysgrifenedig yn anodd. Rhaid rhoi blaenoriaeth i safoni os mai'r nod yw hybu defnydd a bri'r iaith. Mae gan dafodieithoedd eu lle wrth gwrs, yn enwedig mewn peuoedd llafar a chyweiriau anffurfiol.
- Cynnwys digidol: heb sylfaen (llyfrau ysgolheigaidd, cyfieithiadau a Wikipedia yn bennaf) bydd yn amhosibl cynhyrchu offer defnyddiol i'r gymuned ieithyddol.
- Cynnwys agored a meddalwedd cod agored: mae'r penderfyniad i hyrwyddo cynhyrchu cynnwys agored ac i ddefnyddio offer cod agored yn strategaeth hollbwysig er mwyn sicrhau bod y dechnoleg hon yn cael ei datblygu'n raddol ac yn gynaliadwy.

Dadleua Kornai [3] fod y perygl o farwolaeth digidol iaith yn cael ei danamcangyfrif, ac mae'n dod i'r casgliad y bydd llai na 5% o ieithoedd y byd yn gallu goroesi ar ffurf ddigidol. Mae'n dosbarthu ieithoedd i bedwar categori: ieithoedd digidol-ffyniannus, ieithoedd bywiog, ieithoedd treftadaeth ac ieithoedd llonydd.

Gan gadw'r elfennau angenrheidiol hyn mewn cof, hoffwn gyflwyno'r cysyniad BLARK a'r modd y gellir ei addasu i faint y gymuned dan sylw. Mae BLARK ar gyfer iaith [4] yn cynrychioli'r set leiaf o adnoddau sylfaenol (modiwlau meddalwedd, corpora, geiriaduron ac ati) sy'n angenrheidiol i wneud gwaith ymchwil ac i ddatblygu technoleg iaith.

Mae'r papur hwn wedi'i drefnu fel a ganlyn. Ar ôl trafod perthnasedd sawl elfen megis rôl y gymuned ieithyddol, lefel y safoni a faint o destunau sydd ar gael (Adran 2), byddwn yn cyflwyno gwaith cysylltiedig (Adran 3). Yn Adran 4 byddwn yn cyflwyno adnoddau a rhaglenni allweddol i'w rhoi ar waith mewn map ffordd ar gyfer ieithoedd llai eu hadnoddau, gan gynnwys casglu corpora, geiriadur digidol, gwirydd sillafu, morffoleg, anodi corpws, tagiwr rhannau ymadrodd a chloddio testun. Byddwn hefyd yn cyflwyno project cyfieithu peirianyddol gyda'r amcan o'i gwneud hi'n haws defnyddio cofnodion iechyd yng Ngwlad y Basg. Yn olaf, yn Adran 5, byddwn yn trafod ein casgliadau.

2 PERTHASEDD CYNNWYS CYMUNEDOL, SAFONI A DIGIDOL

Mae safoni iaith yn hanfodol i'w llwyddiant fel iaith ysgrifenedig.²

Mae tua 800,000 o siaradwyr a chwe thafodiaith yng Ngwlad y Basg. Mae'r tafodieithoedd yn wahanol iawn i'w gilydd. Yn 1968 penderfynodd Academi'r Iaith Fasgeg greu Basgeg safonol. Ar ôl rhai blynyddoedd o drafodaethau, cafodd ei derbyn yn eang, ac erbyn hyn dyma'r safon (a elwir yn '*Batua*'), a'r model iaith sy'n cael ei defnyddio bron ym mhob testun ffurfiol: yn yr ysgol, yn y brifysgol, wrth weinyddu, ar y rhyngwrwyd, ac ati. Yn y Fasgeg safonol hon y mae newyddiadurwyr teledu a radio ac academyddion yn siarad.

Fel y dywed Hualde a Zuazo [5] "*By any criterion that we may choose, the standardization of Basque in recent years has been a very successful project. Nowadays, standard Basque, which was not developed until the late 1960s, is used in education at all levels, from elementary school to the university, on television and radio, and in the vast majority of all written production in Basque. This success in the societal acceptance of standard Basque is most remarkable given the fact that there is no administration common to all territories where Basque is spoken (divided as they are between Spain and France and even, within Spain, into two separate administrative regions with different legislation regarding the Basque language) and that Basque speakers are almost always fully bilingual in either Spanish or French, so that the existence of a standard Basque language is not strictly required for communication beyond the local level.*"

Rydym am bwysleisio perthnasedd y gwaith a wnaed gan y gymuned ieithyddol yn y broses hon; y gymuned a bwysodd am benderfyniad academiaidd a gwleidyddol i dderbyn y safon, a'r gymuned a gynhyrchodd adnoddau newydd gan ddefnyddio'r safon (llyfrau, cylchgronau, geiriaduron, papur newydd, Wikipedia...). Mae wedi chwarae rôl arbennig o bwysig i ysgolion Gwlad y Basg (*Ikastolak*) o ran adfer a safoni'r Fasgeg [6].

Yr oedd yn bwysig iawn i ni fod y Fasgeg safonol wedi'i diffinio a'i derbyn yn eang cyn i'n grŵp ymchwil ddechrau datblygu offer neu raglenni NLP newydd. Pan oedd angen gwybodaeth ieithyddol arnom, doedd dim angen i ni ei chreu ein hunain, gan fod y gwaith hwn wedi ei wneud yn barod. Doedd dim angen i ni ddelio ag amrywiadau tafodieithol gwahanol am air nac angen dewis ymhlith yr amrywiadau hynny, gan fod Academi'r Fasgeg (*Euskaltzaindia*) eisoes wedi gwneud hynny. Nawr bod angen corpora arnom ar gyfer dysgu neu gasglu, mae'n haws dod o hyd i destunau digonol. Mae'r agwedd hon wedi dod yn ffactor allweddol yn ein llwyddiant oherwydd corpora testun yw'r deunydd crai ar gyfer y prif baradeim technolegol cyfredol: peirianeg iaith sy'n seiliedig ar ddata.

Erbyn hyn, wrth gydweithio ag academyddion neu gymunedau eraill i ddatblygu technoleg ar gyfer ieithoedd llai eu hadnoddau, rydym yn ymwybodol iawn o bwysigrwydd safoni iaith. Er enghraifft, mae uno'r ymdrechion dros Cethswa wedi bod yn dasg anodd oherwydd bod siaradwyr ym Mheriw, Bolifia ac Ecwador yn defnyddio amrywiolion gwahanol o'r un iaith.³

Fel y byddwn ni'n esbonio isod, mae Wikipedia wedi dod yn adnodd allweddol, nid yn unig fel corpws testun ond fel sail addas ar gyfer datblygu offer a rhaglenni newydd. Yn anffodus, weithiau does dim cytundeb ymysg cymunedau lleol i ddiffinio a hyrwyddo amrywiolyn safonol o'r iaith. Y canlyniad i hyn fel arfer yw Wikipedia llai, defnydd aneffeithlon o adnoddau dynol, a chymuned sy'n fwy rhanedig. Mae'r cwestiwn a ddylid defnyddio Nahuatl clasurol neu beidio, er enghraifft, yn dal i fod yn un dadleuol.⁴

²Gellid dadlau nad yw'n bwysig yn achos prosesu llais, ond mae'r rhan fwyaf o systemau llais-i-destun angen adnoddau sy'n seiliedig ar destunau safonol.

³Rydym ni'n gwybod bod modd ystyried yr amrywiolion fel ieithoedd gwahanol, ac yma mae'n rhaid i gymunedau benderfynu a yw'n well ganddyn nhw ymdrechion ar y cyd neu weithio ar wahân.

⁴https://meta.wikimedia.org/wiki/Proposals_for_closing_projects/Closure_of_Classical_Nahuatl_Wikipedia

Wrth gwrs, mae tafodieithoedd ac amrywiolion ieithyddol hefyd yn bwysig i'r gymuned ieithyddol⁵, ond yn ein barn ni mae'r iaith safonol yn flaenoriaeth o ran prosesu testun.

3 BLARK A CHOD AGORED

Cynigiodd Krauwer [4] y "*Basic Language Resource Kit* (BLARK)" fel map ffordd o offer i'w datblygu ar gyfer pob iaith gan ddefnyddio'r derminoleg a ddiffiniwyd mewn menter ar y cyd rhwng ELSNET (*European Network of Excellence in Language and Speech*) ac ELRA (*European Language Resources Association*) yn 1998. Ym mhob un o'r rhain ceir rhestr o adnoddau ac offer sylfaenol. Mae'r term BLARK wedi bod yn llwyddiannus iawn ac fe'i defnyddir mewn nifer fawr o bapurau yn y maes. Yn ddiweddar, mae Mager et al. [7] wedi disgrifio heriau technolegau iaith i ieithoedd cynhenid America ac mae Hassani [8] wedi cynnig dylunio pecyn cymorth BLARK ar gyfer Cwrdeg, sydd â nifer fawr o dafodieithoedd.

Mae Streiter et al. [9] wedi adrodd ar broiectau HLT ar gyfer ieithoedd anghanol ac wedi cynnig canllawiau i gyrff cyllido a strategaethau ar gyfer datblygwyr. Defnyddiant y term "*non-central*" (anghanolog) ac maent yn pwysleisio pwysigrwydd defnyddio meddalwedd am ddim i wella canlyniadau'r gwaith. Mae'r bennod am y manteision a'r problemau sydd heb eu datrys wrth ddefnyddio meddalwedd cod agored gydag ieithoedd anghanol yn ddiddorol iawn. Dadleua Forcada [10] ei bod hi'n bosib defnyddio cyfieithu peiranyddol cod agored ar gyfer ieithoedd bach.

Mae rhwydwaith rhagoriaeth ELSNET wedi paratoi diffiniadau ar gyfer "map ffordd adnoddau iaith a gwerthuso". Dosbarthir yr elfennau yn y diagram (cynnyrch HLT) yn dair is-set: Adnoddau Iaith, Prosesu Iaith a Defnydd Iaith. (Yn ein fersiwn ni mae'r rhain yn ymddangos fel Adnoddau Iaith, Offer Iaith a Chymwysadau Iaith).

Ar sail sawl dangosydd rydym wedi cynnig chwe lefel i gategoreiddio lefel addasiad ieithoedd i dechnoleg [1]:

1. Saesneg: Mae tua 45% o'r holl dudalennau gwe yn y byd wedi'u hysgrifennu yn Saesneg. Mae bron pob rhaglen HLT ar gael yn Saesneg. Gwneir y rhan fwyaf o ymchwil drwy brofi testunau Saesneg.
2. Y 10 prif iaith nesaf sy'n gwasanaethu bron i 50% o ddefnyddwyr y Rhynggrwyd: Dyma'r ieithoedd y mae datblygu adnoddau'n parhau'n weithredol ar eu cyfer ac mae'r rhan fwyaf o gwmnïau mawr ar y rhynggrwyd yn eu cefnogi. Mae Streiter et al. (2006) yn eu galw'n ieithoedd canolog.
3. Tua 70 o ieithoedd sydd ag unrhyw adnoddau HLT o gwbl wedi'u cofrestru. Weithiau gelwir y rhain yn ieithoedd anghanol.
4. Tua 300 o ieithoedd sydd ag unrhyw adnoddau geiriadurol ar-lein wedi'u cofrestru yn yourdictionary.com. Mae'r rhain yn cyfateb bron yn union i'r set o ieithoedd sydd ar Wikipedia neu'r set o ieithoedd sydd wedi diffinio eu safon. Defnyddir y term iaith lai ei hadnoddau (neu brin ei hadnoddau) i ddisgrifio'r ieithoedd hyn (a'r lefel flaenorol hefyd).
5. Tua 2,000 o ieithoedd, sef 2,014 o ieithoedd sydd â systemau ysgrifennu [2].
6. Dyma'r categori mwyaf, gan gynnwys ieithoedd sydd ar gael ar lafar yn unig (mwy na 4,000). Gall y rhan fwyaf ohonynt gael eu hystyried yn ieithoedd mewn perygl.

Yn yr adran nesaf byddwn yn ceisio diffinio, ar sail ein profiad ni, yr adnoddau a'r rhaglenni pwysicaf i'w datblygu fel map ffordd ar gyfer yr ieithoedd yng nghategoriâu 4 a 5. Gall yr ieithoedd hyn gyd-fynd â'r rhai sy'n cael eu hystyried yn '*fywlog*' gan Kornai [3].

⁵Mae rhwydweithiau cymdeithasol (yn enwedig Twitter) wedi dod yn adnodd pwysig ar gyfer adnabod neu drin amrywiadau a thafodieithoedd

Yn ogystal â hyn, rydym am bwysleisio sut y gall y cymunedau ieithyddol ac academiaidd gydweithredu wrth eu datblygu. Mewn rhai achosion, e.e. trychinebau naturiol, gall ymateb yn gyflym fod yn bwysig [11], ond yn gyffredinol mae'n well gosod cynllun sy'n seiliedig ar sefyllfa benodol yr iaith: nifer y siaradwyr, eu cysylltedd, yr adnoddau digidol sydd ar gael, lefel yr integreiddio mewn addysg, ac ati.

Os oes grŵp pwysig o ddefnyddwyr rhyngwryd, mae datblygu offer cydweithredol yn llwybr cynhyrchiol iawn. Offer Wikimedia (Wiktionary, Wikipedia, ac ati) yw'r rhai mwyaf adnabyddus, ond mae adnoddau eraill, fel y llwyfannau torfoli [12] y gellir eu defnyddio gan gymunedau ieithyddol. Pan fydd defnyddwyr rhyngwryd yn brin, mae dod o hyd i gydweithwyr yn y byd academiaidd ac mewn ysgolion yn fwy addas.

4 ADNODDAU A RHAGLENNI ALLWEDDOL

Yn yr isadrannau nesaf byddwn yn cynnig map ffordd cadarn ar gyfer ieithoedd prin eu hadnoddau, gan ddechrau o'r adnoddau, offer a rhaglenni mwyaf sylfaenol. Rydym wedi dewis adnoddau ac offer cod agored yn bennaf. Mae'r map hwn yn seiliedig ar ein profiad ni ac ar gynnydd Streiter et al. [9].

Ni fyddwn yn cynnwys cyfieithu peiranyddol ymysg y rhaglenni gan fod angen mwy o adnoddau ar y dechnoleg hon na sydd ar gael i ieithoedd yn yr ystod 4 i 5. Beth bynnag, os oes gan ieithoedd sy'n perthyn yn agos fwy o adnoddau, gellir llunio peiriant cyfieithu ar sail ieithoedd tebyg heb fawr o ymdrech. Mae Apertium⁶ [10] yn enghraifft dda o'r dechneg hon.

Mae'n bwysig tanlinellu nad yw'r adnoddau rydyn ni'n eu cynnig yn seiliedig ar ddysgu peiranyddol na dysgu dwfn gan fod angen rhagor o adnoddau cyfrifiadurol a data ar gyfer y dulliau hyn. Fodd bynnag, yn ystod y blynyddoedd diwethaf, gwnaed datblygiadau pwysig o ran cymhwyso technegau niwral i ieithoedd sy'n brin o ddata.

Rydyn ni o'r farn bod angen cymryd camau i ddatblygu technolegau iaith sylfaenol cyn symud ymlaen i arbrofion niwral neu ystadegol.

5 CASGLU CORPWS A GEIRIADUR DIGIDOL

5.1 Corpws.

Corpws uniaith yw'r adnodd sylfaenol cyntaf ar gyfer technoleg iaith. Y nodwedd bwysicaf yw ei faint, ond mae nodweddion eraill i'w hystyried hefyd: normaleiddio/amrywiolion, parthau, ffynonellau unigol/lluosog, ac ati. Gall fod yn broiect mawr os ydym am lunio "corpws cenedlaethol" neu "gorpws monitro", gan ddefnyddio metadata (XML/TEI yw'r dull safonol ar gyfer hyn) ac offer ychwanegol.

Mae Wikipedia yn ddewis da ar gyfer echdynnu corpora, ond mewn achosion lle nad yw Wikipedia'n bodoli ar gyfer yr iaith neu os yw'n rhy fach, gall defnyddio technegau "gwe fel corpws" fod yn ddewis da os oes digon o destunau ar gael ar y we. Os nad oes digon o destunau ar gael drwy'r we, sganio testunau neu gydweithio â golygyddion, athrawon ac academyddion yw'r dewisiadau eraill posib.

Disgrifiwyd technegau "gwe fel corpws" gan Kilgarraiff a Grefenstette [13], ac mae Webcorp⁷ yn offeryn defnyddiol at y nod hwn. Weithiau mae angen addasu'r rhaglen i allu delio â nodweddion ieithyddol sy'n benodol i'r iaith [14].

⁶<https://www.apertium.org>

⁷<http://www.webcorp.org.uk/live/>

Pan fo angen sganio dogfennau neu gasglu ffeiliau digidol, mae'n bwysig rhagweld a mesur hyd a lled y gwaith: casglu dogfennau neu ffeiliau mewn gwahanol fformatau, delio â thrwyddedau a materion cyfreithiol, sganio neu drosi fformatau, OCR, mewnosod metadata, ac yn y blaen. O'n profiad ni [15] mae hon yn dasg fawr, llawer mwy na'r disgwyl. Mae Gutierrez-Vasques et al. [16] yn cynnig enghraifft o gasglu corpws dwyieithog.

Ceir projectau byd-eang ar gyfer llunio corpora ar gyfer nifer o ieithoedd [17], [18].

Gellir datblygu'r rhaglenni cyntaf ar sail y corpws a gasglwyd. Gall y rhain gynnwys: enghreifftiau ar gyfer dysgu'r iaith, geiriadur amllder geiriau, gemau sylfaenol (chwilio am eiriau byr, geiriau hir a phalindromau...). Mae'r *Natural Language Toolkit* (NLTK)⁸ yn adnodd defnyddiol iawn ar gyfer datblygu rhaglenni o'r fath.

5.2 Geiriadur Digidol.

Mae geiriadur yn adnodd allweddol i fyfyrwyr, ac yn adnodd pwysig iawn yn gyffredinol. O'n profiad ni, ynghyd â'r gwirydd sillafu, dyma'r adnodd mwyaf ymarferol yr ydym wedi'i ddatblygu. Os yw ar gael, gall *Wiktionary* fod yn sylfaen i'r geiriadur, ond gellir hefyd ei ddatblygu ar sail corpws neu eiriadur sy'n bodoli eisoes⁹.

Gall corpws fod yn ddefnyddiol ar gyfer profi ansawdd a dod o hyd i gofnodion newydd, ond yr opsiwn gorau yw gwaith geiriadurol blaenorol. Rydym yn gwybod o brofiad nad oes modd cael gafael ar eiriaduron rhai cymunedau drwy'r rhyngwrwyd, neu weithiau gallent fod dan drwydded berchnogol. Mae trosi'r geiriadur hwn yn eiriadur ar-lein amlgyfrwng ar sail cronfa ddata eiriadurol yn dasg hollbwysig.

Un profiad da i ni oedd trawsffurfiad lled-awtomatig "*Diccionario Básico Escolar*"¹⁰ Sbaeneg Ciwba [19]. Ar gyfer y Fasgeg, mae Euskalbar¹¹ (ategyn ar gyfer porwyr gwe sy'n anfon ymholiadau cydamserol at eiriaduron a chorpora ar-lein, ac sy'n dangos yr holl ganlyniadau ar yr un pryd) yn adnodd allweddol i'r gymuned.

Gellir defnyddio'r geiriadur i ddatblygu rhaglenni newydd, yn enwedig ar gyfer myfyrwyr. I'r perwyl hwnnw, roedd ein grŵp ni'n rhan o'r gwaith o ddatblygu'r fersiwn Basgeg o *Apalabrados*.¹²

6 SILLAFU A MORFFOLEG

Fel y dywedon ni uchod, mae gwirwyr sillafu'n tueddu i fod ymysg y rhaglenni mwyaf llwyddiannus ar gyfer iaith. Bydd myfyrwyr, athrawon, newyddiadurwyr, awduron a mwy i gyd yn eu defnyddio. Mae hyd yn oed yn fwy angenrheidiol pan fo'r system ysgrifenedig ar gyfer iaith yn cael ei datblygu. Ymhellach, yn achos y Fasgeg, bu'n arf effeithiol iawn yn y broses safoni.

Gall gwiriwr sillafu gael ei greu o gorpws mawr (da), ond bydd ei ansawdd a'i gysondeb yn well os yw'n cael ei ddatblygu ar sail ddadansoddwr morffolegol. Mae'r dull olaf hwn yn hanfodol ar gyfer ieithoedd morffolegol gyfoethog.

⁸<http://www.nltk.org/>

⁹Mae [yourdictionary.com](http://www.yourdictionary.com/) yn cynnig dolenni i adnoddau geiriadurol ar-lein (<http://www.yourdictionary.com/languages.html>) ar gyfer 307 o ieithoedd.

¹⁰<http://ixa2.si.ehu.es/dbe>

¹¹<https://addons.mozilla.org/eu/firefox/addon/euskalbar/>

¹²<http://www.apalabrados.com/>

Gall dadansoddydd morffolegol ddarganfod holl segmentiadau morffolegol posibl gair, y lemas sy'n ei ddefnyddio, a chategori ei ran ymadrodd. Ar sail y wybodaeth hon, gall y gwirydd sillafu benderfynu a yw geiriau sydd heb ddadansoddiad morffolegol yn gamgymeriadau neu'n amrywiolion.

Er mwyn adeiladu'r dadansoddydd, mae angen nodi: (1) y set o lemas gyda'u categorïau, (2) y dodiadau, (3) y morffotactegau sy'n disgrifio cysylltiadau dilys rhwng lemas a dodiadau a'r (4) newidiadau morffoffonolegol a gynhyrchir wrth gysylltu lemas a dodiadau. Gellir cael gwybodaeth am y cam cyntaf, y set o lemas, o'r geiriadur digidol, a'r gweddill gan academyddion neu o lyfrau gramadeg sylfaenol. Mae adnoddau ar gael i gyfuno'r elfennau hyn i gyd. Yn ein hachos ni, defnyddiwyd y ddau adnodd mwyaf poblogaidd: *foma*¹³ a *hunspell*¹⁴. Mae gan y cyntaf [20] gymhelliant ieithyddol o ansawdd uwch ac mae'n symlach ar gyfer disgrifio. Ond mae defnyddio'r ail un wedi bod yn fwy llwyddiannus oherwydd bod modd integreiddio'r disgrifiad yn uniongyrchol i nifer o becynnau meddalwedd (Libreoffice, Mozilla...) ¹⁵. Ar gyfer y Fasgeg [21] ac ar gyfer Cetswha [22] cyfunwyd y ddau opsiwn, gan greu'r disgrifiad yn gyntaf gan ddefnyddio *foma* ac wedyn ei drosi i *hunspell* yn awtomatig¹⁶.

Wrth gwrs, mae gan y gymuned ran bwysig i'w chwarae yn y gwaith o ddatblygu a dosbarthu'r gwirydd sillafu: ei brofi, ei hybu, helpu defnyddwyr newydd i'w osod ar eu cyfrifiaduron, anfon adborth ar gamgymeriadau neu lemas sydd ar goll, ac ati.

7 ANODI, TAGIO RHANNAU YMADRODD A CHLODDIO TESTUN

Mae corpora testun amrwd yn adnoddau hwylus ar gyfer datblygu rhaglenni NLP sylfaenol iawn, ond mae corpora sydd wedi'u hanodi â gwybodaeth forffolegol, gystrawennol neu semantig yn agor y drws i ddatblygu tagwyr rhannau ymadrodd ac offer cloddio testun.

Er enghraifft, fe wnaethon ni lunio EPEC (Corpws Cyfeirio ar gyfer Prosesu'r Fasgeg) ar gyfer y Fasgeg¹⁷ i fod yn gorpws 300,000 gair mewn Basgeg ysgrifenedig safonol. Cafodd ei dagio â llaw ar wahanol lefelau, gan gynnwys morffogystrawen, ymadroddion cystrawennol, a mwy. Mae eisoes wedi cael ei ddefnyddio ar gyfer datblygu adnoddau fel tagiwr rhannau ymadrodd.

Mae'r tagiwr rhannau ymadrodd yn offeryn allweddol arall ynghyd â'r geiriadur digidol a'r gwirydd sillafu, gan ei fod yn gam pwysig o ran paratoi ar gyfer cloddio testun. Mae'n galluogi echdynnu ffeithiau, adnabod endidau (pobl, llefydd, sefydliadau), echdynnu terminoleg, symleiddio testun, a mwy. Mae'r tagiwr yn neilltuo rhan ymadrodd i bob gair yn y testun ar sail ei ddiffiniad (neu ddadansoddiad morffolegol) a'i gyd-destun.

Rydym eisoes wedi trafod y modd y gellir datblygu nifer o raglenni cloddio testun ar sail tagio rhannau ymadrodd, ond gellir ei ddefnyddio i greu offer mwy pwerus hefyd. Mae'r fframwaith piblinellau IXA [23] yn enghraifft o sut i adeiladu'r offer newydd hyn yn hawdd. Mae'n set fodiwlaid o adnoddau (neu biblinellau) Prosesu Iaith Naturiol sy'n darparu mynediad hawdd at dechnoleg NLP ar gyfer nifer o ieithoedd. Mae'n cynnig anodi ieithyddol cadarn

¹³<https://fomafst.github.io/>

¹⁴<http://hunspell.github.io/>

¹⁵<https://addons.mozilla.org/en-US/firefox/language-tools/> Rhestr o'r gwirwyr sillafu a gefnogir gan Mozilla.

¹⁶Mater arall yw mai Microsoft Office yw'r adnodd sy'n cael ei ddefnyddio fwyaf gan ddefnyddwyr. Mae Streiter et al. (2006) yn trafod hyn

¹⁷Arweiniodd ein gwaith ar safoni adnoddau i ni fabwysiadu safonau TEI ac XML fel sail ar gyfer anodi ieithyddol (Artola et al., 2009).

ac effeithlon, sy'n ddefnyddiol iawn o ran cloddio testun. Mae'n hawdd addasu'r dechnoleg agored hon ar gyfer unrhyw iaith arall, a'r unig beth sy'n angenrheidiol yw mynediad at gorpws sydd wedi'i anodi'n ieithyddol.

8 Y FASGEG, NLP A'R MAES CLINIGOL

Mae defnydd offer cyfieithu peirianyddol wedi dod yn rhan annatod o fywyd bob dydd ein cymdeithas. Yn 2019 cafodd ein grŵp IXA y cyfle i gydweithio ag Osakidetza (sefydliad iechyd swyddogol Gwlad y Basg) i ddatblygu adnodd a addaswyd i'r maes clinigol drwy fanteisio ar yr amgylchiadau technolegol newydd sydd ohoni (hynny yw, paradeim llwyddiannus rhwydweithiau niwral mewn cyfieithu peirianyddol) yn ogystal â'r amodau proffesiynol newydd (y cynnydd yn niferoedd y staff dwyieithog sy'n dymuno gweithio drwy gyfrwng y Fasgeg a'r nifer sylweddol o feddygon ifanc newydd sydd wedi'u hyfforddi drwy gyfrwng y Fasgeg yn y brifysgol).

Roeddem wedi bod yn gweithio ar gyfieithu terminoleg CT SNOMED i'r Fasgeg [24], [25] ac ar gyfieithu cofnodion clinigol gan ddefnyddio cyfieithu peirianyddol niwral Basgeg [26], [27].

Nid cyfieithu na datblygu peiriannau cyfieithu yw amcan terfynol cynlluniau swyddogol Gwlad y Basg, dim ond adnodd ychwanegol y byddai'n ddefnyddiol ei gael ydyw. Gwir amcan cynlluniau swyddogol Gwlad y Basg, yn ogystal ag Osakidetza, yw cynyddu presenoldeb a defnydd y Fasgeg mewn cofnodion clinigol, a rhaid dangos a fydd yr adnodd yn cyfrannu at y nod hwn. Yn wir, lansiwyd Itzulbide fel project ymchwil ar sail y rhagdybiaeth os bydd modd dysgu'r system cyfieithu peirianyddol parth cyffredinol i gyfieithu yn y maes clinigol, yn y dyfodol bydd gennym offeryn cyfieithu cyflym a dibynadwy. Cawn weld ymhén ychydig o flynyddoedd a fydd y rhagdybiaeth hon yn cael ei phrofi'n gywir ai peidio.

Dechreuodd y project Itzulbide ym mis Mehefin 2019. Mae'r hyrwyddwyr (Grŵp IXA yr UPV/EHU a'r gweithgor Osakidetza Itzulbide) wedi cynnal cyflwyniadau agored ar y project i drafod safbwyntiau ac amheuron arbenigwyr ac i gasglu adborth. Ar adeg ysgrifennu'r testun hwn, roedd 68 o weithwyr proffesiynol o wahanol arbenigeddau a chefnidiroedd yn gweithio ar greu testunau clinigol dwyieithog i'r project. Dyma ddiolch ac anfon ein hanogaeth i'r holl gyfranogwyr!

Nid yw'r project cyfieithu peirianyddol "Itzulbide" yn atal nac yn cyfyngu ar yr amcanion iaith a'r mesurau normaleiddio cyflenwol a phenodol eraill a gynhwysir yng Nghynllun Iaith Basgeg Osakidetza ar hyn o bryd.

Os caiff ei ddefnyddioldeb ei brofi, bydd yr adnodd yn cael ei integreiddio i system wybodaeth Osakidetza. Gallai hefyd ymestyn i'r gymuned gofal iechyd gyfan (gweithwyr proffesiynol mewn cwmnïau cyhoeddus a phreifat, fferyllwyr, myfyrwyr a darlithwyr prifysgol, peuoedd iechyd nad ydynt yn academaidd, preswylwyr, ysbytai a chymdeithasau proffesiynol) ac i bob man lle siaredir y Fasgeg. Gall hefyd fod yn adnodd cymorth i weithwyr proffesiynol sy'n dysgu'r Fasgeg. I grynhoi, gallai'r defnydd posibl o Itzulbide fynd y tu hwnt i gyfieithu cofnodion clinigol.

Gall project o'r math hwn greu amheuron, ond byddwn yn profi ac yn mesur a yw'r adnodd hwn yn ein dod â ni'n nes at amcanion Cynllun Iaith Gwlad y Basg.

9 NEWYDDION DA AM DDYSGU PEIRIANYDDOL, DYSGU DWFN A MODELAU IAITH BERT

Nid yw dysgu peirianyddol, dysgu dwfn a modelau iaith BERT o reidrwydd yn ddefnyddiol i ieithoedd prin eu hadnoddau. Er mwyn defnyddio'r dulliau hyn, mae angen llawer iawn o ddata ac adnoddau cyfrifiadurol. Fodd bynnag, yn ystod y blynyddoedd diwethaf, gwnaed datblygiadau pwysig o ran cymhwyso technegau niwral i ieithoedd sy'n brin o ddata.

Mae datblygiadau sylweddol wedi bod, hyd yn oed ar gyfer ieithoedd llai eu hadnoddau, mewn sawl maes: echdynnu geirfa [28], anwythiad morffolegol[29], tagio rhannau ymadrodd [30] a chyfieithu peirianyddol [31]. Yn y rhan fwyaf o achosion defnyddir dysgu trawsieithog, ond ceir canlyniadau da hefyd drwy ddefnyddio corpora uniaith (sy'n ddefnyddiol ar gyfer ieithoedd lle nad oes llawer o destunau cyfochrog). Fodd bynnag, dadleua Artetxe et al. [32] fod senarios heb ddata cyfochrog a digonedd o ddata uniaith yn annhebygol o godi mewn gwirionedd.

Profodd modelau BERT amlieithog a Basgeg wedi'u hyfforddi'n barod yn ddefnyddiol iawn yn ddiweddar mewn tasgau NLP ar gyfer y Fasgeg, er eu bod wedi'u creu â chorpora sy'n 500 gwaith yn llai na'r rhai Saesneg a chyda corpws Wikipedia sydd 80 gwaith yn llai¹⁸. Mae mewnbliadau geiriau a modelau iaith parod yn ein galluogi i greu cynrychioliadau cyfoethog o destun ac wedi galluogi gwelliannau ar draws y rhan fwyaf o dasgau NLP. Yn anffodus, maent yn ddrud iawn i'w hyfforddi, ac mae llawer o gwmnïau bach a grwpiau ymchwil yn tueddu i ddefnyddio modelau sydd wedi'u rhag-hyfforddi ac sydd ar gael drwy drydydd partion, yn hytrach na chreu eu rhai eu hunain. Mae hyn yn broblemus, oherwydd yn achos llawer o ieithoedd caiff y modelau eu hyfforddi ar gorpora llai (neu is eu hansawdd). Yn ogystal, nid oes modelau uniaith parod ar gael bob amser i ieithoedd nad ydynt yn Saesneg. Ar y gorau, mae modelau ar gyfer yr ieithoedd hynny wedi'u cynnwys mewn fersiynau amlieithog, lle mae pob iaith yn rhannu'r cwota o is-linyddau a pharmedrau â gweddill yr ieithoedd. Roedd hyn yn arbennig o wir ar gyfer ieithoedd llai fel y Fasgeg. Ond fis Ebrill diwethaf cyflawnodd ein modelau uniaith Basgeg ganlyniadau llawer gwell na'r fersiynau sydd ar gael i'r cyhoedd mewn tasgau NLP sylfaenol, gan gynnwys dosbarthu pynciau, categoreiddio emosiwn, tagio rhannau ymadrodd a NER [33]. Ystyriwch fod model iaith gwreiddiol BERT ar gyfer y Saesneg wedi'i hyfforddi yn 2018 gan ddefnyddio corpws *Google Books*¹⁹ sy'n cynnwys 155 biliwn o eiriau yn Saesneg America, a 34 biliwn o eiriau yn Saesneg Prydain. Mae'r corpws Saesneg bron i 500 gwaith yn fwy na'r un Basgeg. Isod dangosir cyfansoddiad y *Basil Media Corpus* (BMC) a ddefnyddiwyd yn yr arbrawf:

Tabl 1: Cynnwys y Basque Media Corpus (BMC)

Ffynhonnell	Math o destun	Miliynau o docynnau
Wikipedia Basgeg	Gwyddoniadur	35M
Papur newydd Berria	Newyddion	81M
Teledu EiTb	Newyddion	28M
Cylchgrawn Argia	Newyddion	16M
Safleoedd newyddion lleol	Newyddion	224.6M

Ym mis Medi 2020, cafodd IXAmBERT ei ddefnyddio'n llwyddiannus mewn system Ateb Cwestiynau Sgyrsiol Basgeg. Mae IXAmBERT yn fodel iaith amlieithog sydd wedi'i ragosod ar gyfer y Saesneg, y Sbaeneg a'r Fasgeg yn unig. Mae'n cynnwys y corpws Basgeg uniaith ac yn ychwanegu'r Wikipedias Saesneg a Sbaeneg, sy'n cynnwys 2.5 biliwn a 650 miliwn o docynnau yn y drefn honno (80 a 20 gwaith yn fwy na'r Wikipedia Basgeg). Roedd eisoes modd cynnal yr arbrawf trosglwyddo hwn gyda model swyddogol mBERT Google. Ond gan ei fod wedi'i greu i allu delio â llawer iawn o ieithoedd eraill, does dim llawer o gynrychiolaeth i'r Fasgeg ynddo. Y newyddion da yw bod y model hwn wedi cael ei ddefnyddio'n llwyddiannus i drosglwyddo gwybodaeth o Saesneg i'r Fasgeg ar gyfer y

¹⁸<https://www.ehu.eus/ehusfera/ixa/2020/09/30/ixambert-good-news-for-languages-with-few-resources/>

¹⁹<https://www.english-corpora.org/googlebooks/>

system Ateb Cwestiynau Sgyrsiol [34]. Mae'r gwaith a drafodwyd yn y papur yma wedi gosod safon newydd mewn tasgau NLP ar gyfer y Fasgeg, ac mae'r holl feincnodau a'r modelau a ddefnyddiwyd yn y gwaith ar gael i'r cyhoedd²⁰.

10 CASGLIADAU

Mae technoleg iaith yn gymorth pwerus i gymunedau ieithyddol prin eu hadnoddau, er mwyn adfywio'u hiaith a hyrwyddo ei defnydd yn effeithiol.

Ond mae rhai rhagofynion i ddefnyddio Technoleg Iaith. Mae angen cymuned ieithyddol sy'n barod i ddosbarthu a lledaenu adnoddau Technoleg Iaith. Gall sefydlu safon ar gyfer yr iaith a sicrhau ei bod yn cael ei derbyn yn eang hwyluso datblygu offer NLP effeithiol.

Casglu corpora, creu geiriadur digidol, datblygu gwirydd sillafu a tagiwr rhannau ymadrodd, dadansoddi morffoleg, anodi corpora a cloddio testunau yw'r camau cyntaf i'w cymryd. Rydym wedi cyflwyno ein profiad gwerthfawr o ddelio â'r Fasgeg, yn ogystal â rhai awgrymiadau ar gyfer ieithoedd eraill sydd eisiau dylunio map ffordd technoleg iaith.

Cydnabyddiaeth

Mae awduron y gwaith hwn wedi derbyn cefnogaeth ariannol gan y ddau broiect hyn: DOMINO (MINECO-Spain, PGC2018-102041-B-I00 (MCIU/AEI/FEDER, UE)) a TANDO (Gwlad y Basg, ELKARTHEK20/49).

Cyfeiriadau

- [1] Iñaki Alegria, Xabier Artola, Arantza Diaz De Ilarraza a Kepa Sarasola. 2011. Strategies to develop Language Technologies for Less-Resourced Languages based on the case of Basque. Yn Proceedings of 5th Language & Technology Conference: HLT as a Challenge for Computer Science and Linguistics. LTC, Poznan, Gwlad Pwyl, 42-46.
- [2] Lars Borin. 2009. Linguistic diversity in the information society. Yn SALTML2009 Workshop: IR-IE-LRL. Information Retrieval and Information Extraction for Less Resourced Languages. University of the Basque Country, Biscay, Sbaen.
- [3] András Kornai. 2013. Digital language death. PloS one, 8, 10 (Hydref 2013), 1-11.
- [4] Steven Krauwer. 2003. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. Yn Proceedings of SPECOM 2003. SPECOM, Moscow, 8-15.
- [5] José I. Hualde a Koldo Zuazo. 2007. The standardization of the Basque language. Language Problems and Language Planning, 31, 2 (2007), 143-168.
- [6] Irene López-Goñi. 2003. Ikastola in the twentieth century: an alternative for schooling in the Basque Country. History of Education, 32, 6 (2003), 661-676.
- [7] Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra ac Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. Yn Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, UDA, 55-69. [fuzzy]arXiv preprint arXiv:1806.04291.
- [8] Hossein Hassani. 2018. BLARK for multi-dialect languages: towards the Kurdish BLARK. Language Resources and Evaluation, 52, 2 (Mehefin 2018), 625-644.
- [9] Oliver Streiter, Kevin P. Scannell a Mathias Stuflesser. 2006. Implementing NLP projects for non-central languages: instructions for funding bodies, strategies for developers. Machine Translation, 20, 4 (Mawrth 2006), 267-289.
- [10] Mikel L. Forcada. 2006. Open source machine translation: an opportunity for minor languages. Yn Proceedings of the Workshop Strategies for developing machine translation for minority languages LREC (Vol. 6). LREC, Genoa, Yr Eidal, 1-6.
- [11] Robert Munro. 2010. Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. Yn proceedings of the AMTA Workshop on Collaborative Crowdsourcing for Translation. Association for Machine Translation in the Americas, Denver, Colorado, 1-4.
- [12] Marta Sabou, Kalina Bontcheva ac Arno Scharl. 2012. Crowdsourcing research opportunities: lessons from natural language processing. Yn Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies. Association for Computing Machinery, Efrog Newydd, NY, UDA, 1-8.

²⁰ <https://huggingface.co/ixa-ehu/ixambert-base-cased>

- [13] Adam Kilgarriff a Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29, 3 (Medi 2003), 333-347.
- [14] Igor Leturia, Antton Gurrutxaga, Iñaki Alegria, ac Aitzol Ezeiza. 2007. CorpEus, a 'web as corpus' tool designed for the agglutinative nature of Basque. *Yn Building and exploring web corpora*, Proceedings of the 3rd Web as Corpus workshop. Presses universitaires de Louvain, Ffrainc, 69-81.
- [15] Nerea Areta, Antton Gurrutxaga, Igor Leturia, Iñaki Alegria, Xabier Artola, Arantza Diaz de Ilarraza, Nerea Ezeiza ac Aitor Sologaitoa. 2007. ZT Corpus: Annotation and tools for Basque corpora. *Yn Proceedings of Corpus Linguistics 2007*. UCREL, Caerhifryn, 1-19.
- [16] Ximena Gutierrez-Vasques, Gerardo Sierra ac Isaac H.Pompa. 2016. Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl. *Yn Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia, 4210-4214.
- [17] Steven Abney a Steven Bird. 2010. The human language project: building a Universal Corpus of the world's languages. *Yn Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, Uppsala, Sweden, 88-97.
- [18] Kevin P Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. *Yn Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop (4)*. Presses universitaires de Louvain, Louvain-la-Neuve, Gwlad Belg, 5-15.
- [19] Eloíña Miyares Bermúdez, Leonel Ruiz Miyares, Cristina Álamo Suárez, Celia Pérez Marqués, Xabier Artola Zubillaga, Iñaki Alegria Loinaz a Xabier Arregi Iparragirre. 2010. La segunda y tercera ediciones del Diccionario Básico Escolar. *Yn Proceedings of the 14th EURALEX International Congress*. Fryske Akademy, Leeuwarden/Ljouwert, Yr Iseldiroedd, 519-526.
- [20] Mans Hulden. 2009. Foma: a finite-state compiler and library. *Yn Proceedings of the 12th Conference of the EACL: Demonstrations Session*. Association for Computational Linguistics, Athens, Gwlad Groeg, 29-32.
- [21] Iñaki Alegria, Izaskun Etxeberria, Mans Hulden a Montserrat Maritxalar. 2009. Porting Basque morphological grammars to foma, an open-source tool. *Yn International Workshop on Finite-State Methods and Natural Language Processing*. Springer, Berlin 105-113.
- [22] Annette Rios. 2011. Spell checking an agglutinative language: Quechua. *Yn 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. LTC, Poznań, Gwlad Pwyl, 51-55.
- [23] Rodrigo Agerri, Josu Bermudez a German Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. *Yn Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*. LREC, Reykjavik, Gwlad yr Iá, 3823-3828.
- [24] Olatz Perez-de-Viñaspre a Maite Oronoz. 2015. SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. *BMC Medical Informatics and Decision Making*, 15, Suppl 2:S5. doi:10.1186/1472-6947-15-S2-S5
- [25] Olatz Perez-de-Viñaspre a Gorka Labaka. 2016. IXA Biomedical Translation System at WMT16 Biomedical Translation Task. *Yn Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Berlin, Yr Almaen, 477-482.
- [26] Xabier Soto, Olatz Perez-De-Viñaspre, Maite Oronoz a Gorka Labaka. 2019. Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish. *Yn Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*. European Association for Machine Translation, Dulyrn, Iwerddon, 8-18.
- [27] Xabier Soto, Olatz Perez-de-Viñaspre, Gorka Labaka a Maite Oronoz. 2019. Neural Machine Translation of clinical texts between long distance languages. *JAMIA (Journal of the American Medical Informatics Association)*, 26, 12 (Rhagfyr 2019), 1478-1487. <https://doi.org/10.1093/jamia/ocz110>
- [28] Mikel Artetxe, Gorka Labaka ac Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. *Yn Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Yr Eidal, 5002-5007.
- [29] Antonios Anastasopoulos a Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. *Yn Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 984-996.
- [30] Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya ac Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. *Yn Proceedings of the 2017 conference on empirical methods in natural language processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2832-2838.
- [31] Mikel Artetxe, Gorka Labaka, Eneko Agirre a Kyunghyun Cho. 2018. Unsupervised neural machine translation. *Yn Proceedings of the Sixth International Conference on Learning Representations*. ICLR, Vancouver, Canada.
- [32] Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka ac Eneko Agirre. 2020. A Call for More Rigor in Unsupervised Cross lingual Learning. *Yn Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Ar-lein. <https://arxiv.org/pdf/2004.14958.pdf>
- [33] Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa ac Eneko Agirre. 2020. Give your Text Representation Models some Love: the Case for Basque. *Yn Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. LREC, Marseille, Ffrainc, 4781-4788.
- [34] Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa ac Eneko Agirre. 2020. Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque. *Yn Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, Ffrainc, 429-435.

Golwg Agosach ar Fewnblaniadau Geiriau Cymraeg

GERAINT, PALMER

Yr Ysgol Mathemateg, Prifysgol Caerdydd, Cymru

PADRAIG, CORCORAN

Yr Ysgol Cyfrifiadureg a Gwybodeg, Prifysgol Caerdydd, Cymru

LAURA, ARMAN

Yr Ysgol Saesneg, Cyfathrebu ac Athroniaeth, Prifysgol Caerdydd, Cymru

DAWN, KNIGHT

Yr Ysgol Saesneg, Cyfathrebu ac Athroniaeth, Prifysgol Caerdydd, Cymru

IRENA, SPASIC

Yr Ysgol Cyfrifiadureg a Gwybodeg, Prifysgol Caerdydd, Cymru

Mae'r bennod hon yn cyflwyno set o fewnblaniadau geiriau cyffredinol ar gyfer yr iaith Gymraeg. Cynrychiolaeth fector o eiriau mewn iaith yw'r rhain sy'n adlewyrchu eu hystyr semantig mewn ffurf fathemategol, ac maent yn medru cefnogi amryw o dasgau prosesu iaith naturiol, ac yn benodol dulliau dysgu dwfn. Trafodwn ansawdd y corpws hyfforddi a'r mewnblaniadau a gafodd eu dysgu'n awtomatig o'r corpws. Defnyddiwn ddadansoddiad ansoddol er mwyn adnabod unrhyw welliannau potensial i'r mewnblaniadau sy'n arbennig i'r Gymraeg.

Allweddeiriau: mewnblaniadau geiriau, technolegau iaith, yr iaith Gymraeg

1 CYFLWYNIAD

Mae'r bennod hon yn cyflwyno gwaith ar hyfforddiant set o fewnblaniadau geiriau ar gyfer yr iaith Gymraeg, sydd yn gallu cefnogi nifer o dasgau prosesu iaith naturiol. Ein nod yw creu set gynhwysfawr o fewnblaniadau geiriau sy'n cynnwys gwybodaeth ar gystrawen a semanteg yr iaith Gymraeg, er mwyn gwella a helpu datblygu modelau prosesu iaith naturiol ar gyfer y Gymraeg. Yn y bennod hon rhoddwn ychydig o gefndir, trafodwn bwysigrwydd ac ansawdd y corpws hyfforddi, ac edrychwn am gliwiau i ansawdd y mewnblaniadau a grëwn trwy archwilio'u gofod fector. Mae'r math hwn o ddadansoddiad yn bwysig oherwydd gall unrhyw wendidau a biasau yn y mewnblaniadau gael eu hetifeddu gan y technolegau iaith sy'n eu defnyddio.

Mae Adran 2 yn rhoi cefndir damcaniaethol, gydag Adran 2.1 yn disgrifio beth yw mewnblaniadau geiriau, Adran 2.2 yn disgrifio'u defnydd potensial, ac Adran 2.3 yn disgrifio'r broses hyfforddi. Yna mae Adran 3 yn disgrifio'r

gwaith, gydag Adran 3.1 yn disgrifio'r corpws, ac Adran 3.2 yn trafod agweddau o ansawdd y mewnbaniadau geiriau.

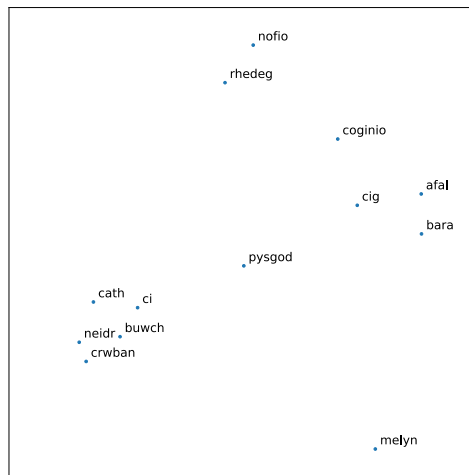
2 CEFNDIR MEWNBLANIADAU GEIRIAU

Mae'r adran hon yn rhoi cefndir damcaniaethol i fewnbaniadau geiriau, beth yw mewnbaniadau geiriau, eu defnydd potensial, a sut y cânt eu hyfforddi.

2.1 Beth yw Mewnbaniadau Geiriau?

Mewnbaniadau geiriau yw mapiad o ofod lecsico-sementig i ofod fector n -dimensiwn gwerth real, $\mathbb{R}^n$ [1]. Hynny yw, mapiad o eiriau mewn iaith i fectorau o rifau real. Gallwn feddwl am hyn fel bod gan bob gair mewn iaith fector real i'w gynrychioli.

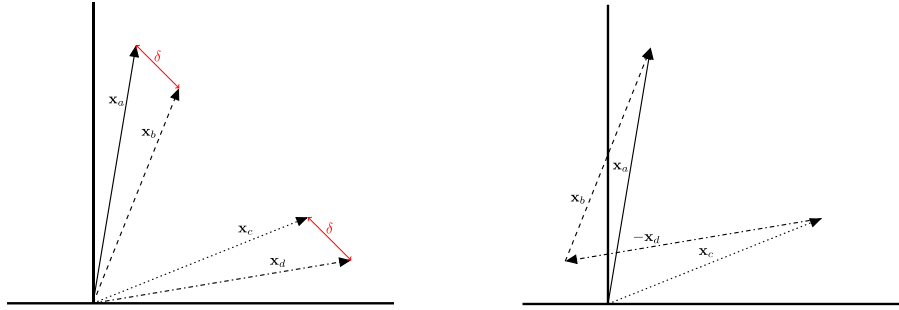
Mae gan fewnbaniadau geiriau o ansawdd da briodweddau penodol yn ymwneud ag ystyr a chyd-destun arferol geiriau. Mewn geiriau eraill, y nod yw i'r fectorau grynhoi gwybodaeth semantig ddosraniadol. Er mwyn i'r mapiad storio gwybodaeth semantig berthnasol, defnyddir fectorau aml-dimensiwn, fel arfer gydag $100 \leq n \leq 500$. Yn ddelfrydol dylai perthnasoedd rhwng fectorau'r mewnbaniadau adlewyrchu'r perthnasoedd semantig rhwng eu geiriau cyfatebol. Os yw $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathbb{R}^n$, lle $\mathbf{x}_1$ yw'r fector sy'n cynrychioli'r gair "llew", $\mathbf{x}_2$ yw'r fector sy'n cynrychioli'r gair "teigr", ac $\mathbf{x}_3$ yw'r fector sy'n cynrychioli'r gair "llyfr", yna fe ddylai $\|\mathbf{x}_1 - \mathbf{x}_2\|$ bod llawer llai nag $\|\mathbf{x}_1 - \mathbf{x}_3\|$, lle'r gweithrediad $\|\cdot\|$ yw rhyw norm yn cynrychioli pellteroedd mewn n -dimensiwn. Hynny yw fe ddylai'r fectorau sy'n cynrychioli "llew" a "teigr" for llawer agosach at ei gilydd na'r fectorau ar sy'n cynrychioli "llew" a "llyfr", oherwydd eu bod yn fwy cysylltiedig. Mae Ffigur 1 yn dangos cynrychioliad o'r syniad hwn mewn dau ddimensiwn, gyda geiriau tebyg yn agosach at ei gilydd na geiriau annhebyg.



Ffigur 1: Enghraifft dau ddimensiwn o berthynas mewnbaniadau geiriau i ystyr semantig.

Priodwedd ddefnyddiol arall o fewnbaniadau geiriau yw bod maint a chyfeiriad y gwahaniaeth rhwng fectorau yn debyg os ac yn unig os yw'r gwahaniaethau semantig rhwng eu geiriau cyfatebol yn fwy neu lai yn hafal. Er enghraifft, os yw $\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c, \mathbf{x}_d \in \mathbb{R}^n$, ac $\mathbf{x}_a$ yw'r fector sy'n cynrychioli'r gair "brenin", $\mathbf{x}_b$ yw'r fector sy'n cynrychioli'r

gair “brenhines”, $\mathbf{x}_c$ yw'r fector sy'n cynrychioli'r gair “athro”, ac $\mathbf{x}_d$ yw'r fector sy'n cynrychioli'r gair “athrawes”; yna fe ddylai teithio o $\mathbf{x}_a$ i $\mathbf{x}_b$ bod yn debyg i deithio o $\mathbf{x}_c$ i $\mathbf{x}_d$, oherwydd bod gan y parau o eiriau “brenin” a “brenhines” ac “athro” ac “athrawes” yr un berthynas â'i gilydd. O briodweddau adio fectorau, mae hyn yn golygu bod $\mathbf{x}_a = \mathbf{x}_c - \mathbf{x}_d + \mathbf{x}_b$, a chaiff hyn ei ddangos yn dau ddimensiwn yn Ffigur 2.



Ffigur 2: Enghraifft of adio fectorau a'u perthynas i ystyron semantig cyfatebol.

Mae'r ffaith y gallwn ni drin geiriau fel gwrthrychau mathemategol, ynghyd â'r priodweddau hollbwysig a ddisgrifir uchod, yn golygu eu bod yn gallu modelu rhesymeg ynglŷn ag iaith naturiol yn defnyddio gweithrediadau fector syml a ddefnyddir yn aml mewn algorithmau dysgu dwfn. Disgrifir rhai cymwysiadau mewnbaniadau geiriau yn yr adran nesaf.

2.2 Defnydd Mewnbaniadau Geiriau

Profwyd bod mewnbaniadau geiriau yn gwella perfformiad nifer o fodelau a thasgau prosesu iaith naturiol. Mae defnyddio set o fectorau, sy'n cynnwys cysylltiadau semantig rhwng geiriau, mewn tasgau prosesu iaith naturiol, yn golygu eu bod yn perfformio'n dda ar y broblem allan-o-eirfa, hynny yw gallant wneud rhagfynegiadau ar eiriau nad ydynt wedi eu gweld i'r blaen (h.y. yn y broses hyfforddi), trwy gymryd mantais ar eu cysylltiadau â geiriau arall yn y data hyfforddi. Mae'r math o dasgau yn cynnwys:

- *Cyfieithu peiranyddol*: mae defnyddio mewnbaniadau geiriau mewn tasgau cyfieithu peiranyddol niwral wedi profi i wella'u perfformiad, er enghraifft yn [2].
- *Dadansoddi sentiment*: tasg yw hon i gategoreiddio brawddegau yn ôl sentiment cadarnhaol, niwtral, neu negyddol. Mae gan hyn gymwysiadau mewn pethau megis ymchwil masnachol a dadansoddi tueddiadau ar y cyfryngau cymdeithasol. Enghraifft lle mae mewnbaniadau geiriau'n gwella perfformiad yw [3].
- *Adnabod endidau*: tasg yw hon sydd â'r nod o adnabod ymadroddion sy'n ymddwyn fel enwau, er enghraifft “prif weinidog” neu “canser y fron”. Mae gan hyn gymwysiadau wrth chwilota cronfeydd data o gorpora, yn enwedig mewn meysydd academiaidd a meddygol. Enghraifft o fewnbaniadau geiriau yn cael ei defnyddio ar gyfer hyn yw [4].
- *Parsio dibyniaethau*: tasg i ddadansoddi strwythur brawddeg trwy labeli geiriau o fewn brawddeg fel rhannau ymadrodd (enwau, berfau, ansoddeiriau ac yn y blaen) ac adnabod y cysylltiadau rhyngddynt. Er enghraifft ([5]) yn y frawddeg “gwelodd Siân y ci” mae'r ferf “gwelodd” yn cysylltu dau enw, y goddrych “Siân” a'r gwrthrych “ci”. Enghraifft o ddefnyddio mewnbaniadau geiriau i wneud hwn yw [6].

Caiff y Gymraeg ei disgrifio fel iaith lai ei hadnoddau yn nhermau prosesu iaith naturiol. Mae hyn yn golygu bod argaeledd data a chorpora yn brin o gymharu â ieithoedd a siaredir yn fwy cyffredin, ac felly gall hyfforddi modelau

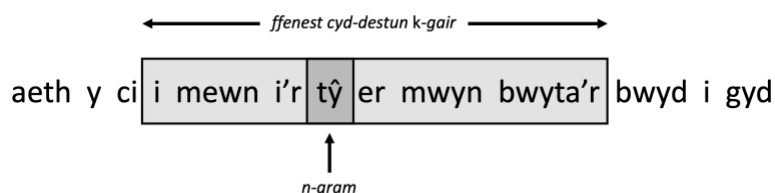
prosesu iaith naturiol sy'n perfformio'n dda fod yn sialens. Felly gall mewnbaniadau geiriau ymddwyn fel galluogwyr cymwysiaidau prosesu iaith naturiol cymhleth.

2.3 Hyfforddi Mewnbaniadau Geiriau

Gelwir unrhyw fapiad o'r gofod geirfaol i ofod fector yn fewnbaniadau eiriau, ond er mwyn iddynt fod yn ddefnyddiol i dasgau prosesu iaith naturiol eraill rydym ni eisiau iddynt arddangos rhai priodweddau penodol, fel y rhai a drafodir yn Adran 2.1. Er mwyn cyflawni hwn, mae nifer o ffyrdd o adeiladu mewnbaniadau eiriau, ac mae technegau dysgu peiranyddol yn dod yn fwyfwy boblogaidd. Yn nhermau dysgu peiranyddol, rydym yn *hyfforddi* mewnbaniadau geiriau o ddata mewnbwn, sef corpws digon mawr o destun.

Yn y gwaith hwn fe wnaethon ni edrych ar ddau ddull o hyfforddi mewnbaniadau geiriau, *Word2Vec* ([7] a *FastText* ([8], sydd yn gallu defnyddio dau algorithm gwahanol, *Skip-gram* ([9] a *Continuous Bag of Words (CBOW)*, [7]). Mae'r rhain yn dibynnu ar ddamcaniaeth semantig ddosrannol, sy'n honni bod geiriau sy'n ymddangos yn yr un cyd-destun yn ceisio cludo ystyron tebyg [1]. Felly mae'r technegau hyn yn tybio fod ystyr gair yn dibynnu'n bennaf ar ei gyd-destun, hynny yw amllder y geiriau sy'n gymdogion agos iddo o fewn brawddegau. Mae *Word2Vec* yn cymryd y syniad hwn yn llythrennol. Ni edrychir ar sillafu gair, caiff pob gair ei drin fel n-gram ar wahân, ac ystyrir yr n-gramau eraill sy'n ymddangos yn agos ato yn fwyaf aml o fewn brawddeg. Mae hyn yn gallu bod yn ddefnyddiol ar gyfer ieithoedd analytig megis Mandarin. Mae *FastText* fodd bynnag hefyd yn ystyried is-eiriau (cydrannau gwahanol o fewn geiriau ysgrifenedig), sy'n ddefnyddiol ar gyfer ieithoedd sydd morffoleg gyfoethog, er enghraifft Tyrceg.

Nod yr algorithmau hyn yw canfod fectorau fel bod cyfanswm y pellter SoftMax rhwng pob n-gram a'i gymdogion yn adlewyrchu'r tebygolrwydd o ganfod gair mewn rhyw gymdogaeth o eiriau eraill. Diffinnir cymdogion n-gram fel yr n-gramau sydd o fewn ffenest cyd-destun iddo mewn brawddeg, hynny yw'r k cymydog agosaf. Dangosir enghraifft yn Ffigur 3. Mae'r algorithm *Skip-gram* yn ceisio adlewyrchu'r tebygolrwydd o'r gair hwnnw'n ymddangos yn amodol ar y cyd-destun, tra bo *CBOW* yn ceisio adlewyrchu'r tebygolrwydd o'r cyd-destun yn ymddangos yn amodol ar y gair hynny. Felly mae eu perfformiadau'n amrywio rhwng y geiriau mwy a llai aml yn y corpws hyfforddi.



Ffigur 3: Enghraifft o air o fewn ei ffenest cyd-destun.

3 MEWNBLANIADAU GEIRIAU CYMRAEG

Fe hyfforddon ni set o fewnbaniadau geiriau ar gyfer yr iaith Gymraeg, fel y'i disgrifir yn [11]. Fe gasglon ni gorpws, a ddisgrifir yn Adran 3.1, yna fe hyfforddon ni fewnbaniadau gan ddefnyddio pob cyfuniad o'r technegau a'r algorithmau a ddisgrifir yn Adran 2.3. Fe werthuson ni'r mewnbaniadau hyn gan ddefnyddio amryw o dechnegau meintiol ([12], [13] ac adroddir y canlyniadau yn [11]. Mae Adran 3.2 yn gwerthuso yn ansoddol y set o fewnbaniadau a berfformiodd orau ar y technegau meintiol hyn.

3.1 Casglu Corpws

Fel y disgrifir yn Adran 2.3, mae angen hyfforddi mewnbaniadau geiriau ar gorpws o ddogfennau iaith Gymraeg, ac mae priodweddau'r corpws yn dylanwadu'n fawr ar briodweddau'r mewnbaniadau geiriau. Ar gyfer ieithoedd a siaredir yn gyffredin, megis Saesneg, nid yw argaeledd corpora yn broblem sylweddol. Er enghraifft defnyddir Wikipedia, sydd ar hyn o bryd yn cynnwys dros 3.7 biliwn o eiriau, er mwyn hyfforddi modelau Saesneg. Oherwydd bod gan y Gymraeg lai o siaradwyr a llai o adnoddau, mae'r dasg o gasglu'r fath gorpws ar gyfer y Gymraeg yn fwy heriol.

Mae nifer o gorpora bach, arbenigol i'w cael ar gyfer y Gymraeg, pob un o faint rhy fach i fod yn ddefnyddiol ar ben ei hun. Felly fe gasglon ni gorpws mwy trwy gyfuno nifer o'r adnoddau Cymraeg hyn, yn ogystal â defnyddio offer crafu'r we (*web scraping*), er mwyn cael gafael ar destunau ychwanegol. Ystyriaeth arall oedd amrywiaeth y corpws, gan fod modelau prosesu iaith naturiol sydd wedi'u hyfforddi ar gorpora arbenigol ond yn ddefnyddiol ar gyfer tasgau mewn parthau arbenigol. Oherwydd mai ein tasg oedd creu set o fewnbaniadau geiriau cyffredinol i'w defnyddio mewn tasgau prosesu iaith naturiol cyffredinol, fe geision ni gasglu corpws digon amrywiol. Yn gyfan gwbl mae ein corpws yn cynnwys 92,963,671 gair. Mae Tabl 1 yn dangos ffynonellau a meintiau'r data a ddefnyddiwyd. Fe gymeron ni ofal i sicrhau nad oedd y ffynonellau yn cynnwys yr un testun gwreiddiol, a'i fod yn cynnwys amrywiaeth o gyweiriau iaith a phynciau gwahanol.

Defnyddir y Beibl yn aml fel y lle cyntaf i edrych wrth adeiladu corpws gan ei fod wedi cael ei gyfieithu i nifer fawr o ieithoedd ac mae fel arfer ar gael am ddim ar ffurf electronig, ond mae ganddo anfanteision o ran cynnwys ei iaith hynafol, ffurfiol, ac amrywiaeth mater pwnc cyfyngedig. Mae'r corpws An Crúbadán yn gasgliad mawr o wefannau, blogiau, a thrydar Cymraeg. Derbynion ni is-set ragbaratoawl o gorpws electronig CorCenCC, sy'n cynnwys e-byst a negeseuon testun. Gyda'i gilydd mae'r rhain yn darparu ystod eang o bynciau mewn Cymraeg cyfoes, efallai anffurfiol. Mae Gwerddon yn gyfnodolyn academaidd aml-ddisgybliedig yn y Gymraeg; ac mae project DECHE yn cynnwys gwerslyfrau academaidd yn y Gymraeg. Mae'r corpora hyn yn darparu samplau o ysgrifennu hynod ffurfiol academaidd mewn amrywiaeth o bynciau arbenigol. Mae'r teclyn Google Corpuscrawler yn declyn cod agored er mwyn cael gafael ar gorpora ar gyfer nifer fawr o ieithoedd gwahanol. Wrth redeg y teclyn ar gyfer y Gymraeg, yn bennaf mae'n cyrchu erthyglau BBC Cymru Fyw. I ychwanegu at y ffynonellau parod, fe wnaethon ni hefyd gyrchu gwefannau cylchgrawn a newyddion eraill, Golwg360, O'r Pedwar Gwynt, Barn, a PoblCaerdydd, sy'n darparu enghreifftiau o iaith newyddiadurol. Mae trafodion y Cynulliad Cenedlaethol ar gael, ac mae dau gorpws parod o'r rhain ar gael, o 1999 i 2006, ac yna o 2007 i 2011, er mwyn sicrhau bod iaith lafar ffurfiol o bynciau gwleidyddol wedi'u cynnwys. Mae'r Gronfa Electroneg o Gymraeg yn cynnwys enghreifftiau o ryddiaith ffuglennol a dogfennau gweinyddol sydd hefyd yn ychwanegu at amrywiaeth yr iaith sydd wedi'i chynnwys yn y corpws, ac mae Wicipedia Cymraeg yn cynnwys erthyglau ar amrywiaeth eang o bynciau, rhai cyffredinol ac arbenigol. Hon yw'r ffynhonnell sengl fwyaf sydd ar gael am ddim yn y Gymraeg.

Tabl 1: Ffynonellau'r is-gorpora

Ffynhonnell	Nifer o eiriau
Y Beibl ¹	749,573
Corpws An Crúbadán [14]	22,572,066

¹<http://www.beibl.net/>

Ffynhonnell	Nifer o eiriau
CorCenCC ²	1,875,540
Cronfa Electroneg o Gymraeg [15]	1,046,800
Google Corpuscrawler ³	14,791,835
Gwerddon ⁴	767,677
Project DECHE [16]	2,126,153
Trafodion Cynulliad Cenedlaethol 1999-2006 [17]	11,527,963
Trafodion Cynulliad Cenedlaethol 2007-2011 [18]	8,883,970
Wikipedia Cymraeg ⁵	21,233,177
Gwefannau amrywiol eraill ⁶	7,388,917
Cyfanswm	92,963,671

3.2 Gwerthuso

Fe ddefnyddion ni'r corpws a ddisgrifir yn Adran 3.1 i hyfforddi nifer o fewnblaniadau geiriau gyda phob cyfuniad o'r technegau a ddisgrifir yn Adran 2.3, *FastText* a *Word2Vec*, *Skip-gram* a *CBOW*, yn ogystal â nifer o ffyrdd o docyneiddio'r gorpws. Fe ddefnyddion ni nifer o dechnegau meintiol i werthuso'r mewnblaniadau geiriau. Mae'r set o fewnblaniadau geiriau a berfformiodd y gorau yn feintiol ar gael i'w lawr lwytho fan hyn: <https://datainnovation.cardiff.ac.uk/is/wecy/>

Mae hefyd yn bwysig gwerthuso'r mewnblaniadau geiriau yn ansoddol, hynny yw gyda'r llygaid, er mwyn gweld pa mor dda mae'r mewnblaniadau geiriau yn cynrychioli semanteg geiriau yn annibynnol o'r dechneg gwerthuso awtomatig a ddefnyddir. Un ffordd safonol o wneud hyn yw dewis gair, chwilio am y set o fectorau eraill mwyaf agos at gynrychiolaeth fector y gair hynny yn y gofod fector, a gwirio os yw'r geiriau sy'n cael eu cynrychioli gan y fectorau hynny yn gysylltiedig yn semantig gyda'r gair gwreiddiol. Mae hyn yn debyg i'r dechneg awtomatig a ddisgrifir yn [19]. Y fantais o wneud hyn yn awtomatig yw gallu archwilio mwy o eiriau yn fwy effeithlon, ond y fantais o wneud hyn gyda'r llygaid yw gallu arsylwi cysylltiadau, nodweddion tebyg, a phatrymau na chafodd eu pennu cyn dechrau. Mae hefyd yn rhoi'r cyfle i ddadansoddi natur y cysylltiadau, ac edrych ymhellach ar y geiriau nad ydynt yn gysylltiedig, sy'n gallu rhoi syniad i ni o welliannau potensial y model.

Fan hyn rydyn ni wedi dewis pum gair ac wedi edrych ar y deg fector agosaf iddynt gan ddefnyddio'r pellter cosin. Mae Tabl 2 yn rhoi'r geiriau a ddewiswyd a'r deg gair agosaf iddynt. Fe wnawn ni edrych arnynt yn eu tro er mwyn gweld beth maent yn dweud wrthom am ansawdd a strwythur y mewnblaniadau geiriau.

²<https://www.corcenc.org/>

³<https://github.com/google/corpuscrawler>

⁴<http://www.gwerddon.cymru/cy/hafan/>

⁵<https://cy.wikipedia.org/wiki/Hafan>

⁶<https://www.golwyg360.cymru> <https://pedwargwynt.cymru/>, <https://barn.cymru/> and <https://poblaerdydd.com/>

Tabl 2: Deg gair agosaf sampl o eiriau.

Gair	Geiriau agosaf
arian	harian, arian', ariannu, cyllid, bunnau, gyllid, bres, arianu, goffrau, wario
cysgu	gysgu, deffro, llewygu, cysgai, chwtsho, chysgu, dihuo, cerdded, ddihuo, gysgai
blawd	blawdog, flawd, blawr, blawdy, siaradblawd, menyn, fenyn, cyflasyddau, blaw, chwstard
tonyrefail	donyrefail, nhonyrefail, ffosyrefail, maesyrefail, trebanog, tonysguboriau, tonteg, tonypandy, trecynon, trealaw
actores	actor, actoresau, sgriptwraig, digrifwraig, berfformwraig, comediwraig, chantores, gantores, ddigrifwraig, perfformwraig

Yn eu tro:

- **“arian”**: Mae'r deg gair agosaf i gyd yn ymwneud â'r pwnc ariannol. Mae hyn yn awgrymu bod gan y model hwn y briodwedd ddymunol gyntaf a drafodwyd yn Adran 2.1. Sylwch hefyd fod yna dreigladau a chamsillafiadau. Er mwyn i'r mewnbaniadau geiriau fod yn ddefnyddiol ar gyfer tasgau prosesu iaith naturiol, mae angen iddynt ddisgrifio'r math o iaith y bydd y tasgau'n cael eu defnyddio arno, ac felly'r math o iaith mae pobl mewn gwirionedd yn ei ddefnyddio. Mae'r cynnwys o dreigladau a chamsillafiadau felly hefyd yn ddymunol, gan mai dyma sut mae pobl yn defnyddio'r Gymraeg mewn bywyd pob dydd, y dystiolaeth am hyn yw fod y geiriau hyn yn ymddangos yn ddigon aml yn ein corpws i ymddangos yn y mewnbaniadau geiriau.
- **“cysgu”**: Mae'r deg gair agosaf i gyd yn ferfau o wahanol ffurfiau, ac yn fras gysylltiedig. Sylwch ar y gair “chwtsho” fan hyn, sydd ddim yn air Cymraeg go iawn, ond sy'n ymddangos fan hyn gan ei fod yn ddigon niferus yn y corpws.
- **“blawd”**: Mae'r deg gair agosaf yn dangos geiriau a chynhwysion cysylltiedig. Mae tri eithriad, “siaradblawd”, “blawr”, a “blaw”. Mae ymchwil pellach yn dangos taw hashnod ar drydar yw'r cyntaf. Mae i'w weld fod y ddau arall, fodd bynnag, dim ond yn ymddangos gan eu bod wedi eu sillafu yn debyg i'r gair gwreiddiol, sy'n digwydd oherwydd bod yr algorithm *FastText* yn ystyried is-eiriau fel n-gramau. Mae hyn yn anghyfleus, ac yn arwydd o faint corpws rhy fach. Nodwn hefyd nad yw'r gair cyfystyr “fflwr” yn ymddangos, ac efallai oherwydd bod y gair hwn yn ymddangos yn fwy aml yn yr iaith lafar nag mewn iaith ysgrifenedig, a chorpws ysgrifenedig yn bennaf sydd gennym.
- **“tonyrefail”**: Heblaw am dreigladau, mae pob gair yn y deg gair agosaf hefyd yn drefi. Ymhellach, mae pob un yn dref yn ne Cymru, yn agos iawn yn ddaearyddol at Donyrefail. Mae'r ffenomenon hon wedi'i harsylwi gyda nifer o drefi bach a chanolig eu maint yng Nghymru, efallai oherwydd bod sôn am drefi cyfagos yn yr un erthyglau newyddion sy'n rhan o'n corpws.
- **“actores”**: Mae'r deg gair agosaf i gyd yn swyddi i wneud â'r diwydiant adloniant. Heblaw am y ddau air cyntaf, sef ffurfiau gwahanol o'r gair gwreiddiol, mae pob gair yn ei ffurf fenywaidd. Er bod y briodwedd o ddeall ffurfiau gwahanol eiriau yn un ddymunol ar gyfer mewnbaniadau geiriau o ansawdd da, efallai fod hyn yn awgrymu bod y model wedi gwahaniaeth rhwng ffurfiau benywaidd a gwrywaidd y swyddi yn fwy na natur y swyddi eu hunain. Nid yw hyn yn dystiolaeth bod hyn wedi digwydd, ond mae angen nodi, trwy hyfforddi technolegau iaith y dyfodol ar ddata'r gorffennol y gall ystrydebau cymdeithasol gael eu hatgyfnerthu gan y model. Mae hon yn broblem adnabyddus, lle gall mewnbaniadau geiriau dangos biasau hiliol a rhywiaethol oherwydd y data hyfforddi ([20]).

4 CASGLIADAU

Ar y cyfan mae'r gwerthusiad ansoddol o'r mewnbaniadau geiriau hyn yn cytuno gyda'r gwerthusiad meintiol, eu bod yn perfformio'n ddigonol, hynny yw fel y disgwylir iddynt wneud o ran adlewyrchu cystrawen a semanteg iaith.

Mae edrych yn agosach at anomaleddau wedi ein helpu ni i ddeall y fectorau yn well, i gydnabod unrhyw wendidau potensial, i archwilio’u hachosion potensial, ac i archwilio ffyrdd i ni atal y gwendidau hyn rhag cael eu hetifeddu gan y tasgau prosesu iaith naturiol fydd yn eu defnyddio. Mae’r mewnbaniadau geiriau a ddisgrifir yn y bennod hon ar gael yn god agored o’r wefan ganlynol: [https://datainnovation.cardiff.ac.uk/is/wecy/yngghyd â theclyn i ganfod deg gair agosaf unrhyw air yn y gofod fector](https://datainnovation.cardiff.ac.uk/is/wecy/yngghyd%20theclyn%20i%20ganfod%20deg%20gair%20agosaf%20unrhyw%20air%20yn%20y%20gofod%20fector).

CYDNABYDDIADAU

Ariannwyd yr astudiaeth hon gan Lywodraeth Cymru fel rhan o brosiect yn gysylltiedig ag isadeiledd prosesu’r iaith Gymraeg. Hoffem ddiolch i Gareth Morlais, CorCenCC, ac i Kevin Scannell am eu cymorth yn casglu a gwerthuso’r corpwys.

CYFEIRIADAU

- [1] Tom Young, Devamanyu Hazarika, Soujanya Poria, ac Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 13, 3 (Awst 2018), 55-75.
- [2] Ye Qi, Devandra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, a Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? *Yn Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 529-535.
- [3] Maria Giatsoglau, Manolis G. Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, a Konstantinos Ch Chatzissavvas. 2017. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications* 69 (Mawrth 2017), 214-224.
- [4] Yonghui Wu, Jun Xu, Min Jiang, Yaoyun Zhang, a Hua Xu. 2015. A study of neural word embeddings for named entity recognition in clinical text. *Yn AMIA Annual Symposium Proceedings 2015*. AIMA, San Francisco, CA, 1326-1333.
- [5] Dewi Bryn Jones, Delyth Prys, Myfyr Prys, a Gruffudd Prys. 2019. Llawlyfr technolegau iaith. *Coleg Cymraeg Cenedlaethol*. Bangor, Cymru.
- [6] Timothy Dozat, a Christopher D Manning. 2018. More accurate semantic dependency parsing. *Yn Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 2 (Short Papers)*. Association for Computational Linguistics, Melbourne, Awstralia, 484-490.
- [7] Tomas Miklov, Kai Chen, Greg S. Corrado, a Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Yn Proceedings of the International Conference on Learning Representations (2013)*. ICLR, Scottsdale, AZ, UDA.
- [8] Piotr Bojanowski, Edouard Grave, Armand Joulin, a Tomos Mikolov. 2017. Enriching word sectors with subword information. *Transactions of the Association for Computational Linguistics* 5, (2017), 135-146.
- [9] Tomos Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, a Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Yn Neural Information Processing Systems. Annual Conference*. 27th 2013. (4 Vols) *Advances In Neural Information Processing Systems* 26. NIPS, Lake Tahoe, Nevada, UDA, 3136-3144.
- [10] Zellig S Harris. 1954. Distributional structure. *WORD* 10 2-3 (1954), 146-162.
- [11] Pdraig Corcoran, Geraint Palmer, Laura Arman, Dawn Knight, ac Irena Spasic. 2020. Word Embeddings in Welsh. *Yn cael ei adolygu*.
- [12] Tobias Schnabel, Igor Labutov, David Mismo, a Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. *Yn Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portiwgal, 298-307.
- [13] Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.
- [14] Kevin P Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. *Yn Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop (4)*. Presses universitaires de Louvain, Louvain-la-Neuve, Gwlad Belg, 5-15.
- [15] N Ellis, C O’Dochartaigh, W Hicks, M Morgan, a M Laporte. 2001. Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh.
- [16] Delyth Prys, Mared Roberts, a Dewi Bryn Jones. 2014. DECHE and the Welsh national corpus portal. *Yn Proceedings of the First Celtic Language Technology Workshop*. Association for Computational Linguistics and Dublin City University, Dulyn, Iwerddon, 71-75.
- [17] Dafydd Jones, ac Andreas Eisele. 2006. Phrase based statistical machine translation between English and Welsh. *Yn Proceedings of the 5th SALT/MIL Workshop on Minority Languages at the 5th International Conference on Language Resources and Evaluation*. LREC, Genoa, Yr Eidal, 75-77.
- [18] Kevin Donnelly. 2013. Kynulliad3: a corpus of 350,000 aligned Welsh and English sentences from the Third Assembly (2007-2011) of the National Assembly for Wales. <http://cymraeg.org.uk/kynulliad3>
- [19] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, ac Eytan Ruppín. 2002. Placing search in context:

the concept revisited. *ACM Transactions on Information Systems* 20, 1 (Ionawr 2002), 116-131.

- [20] James Zou, a Londa Schiebinger. 2018. AI can be sexist and racist - it's time to make it fair. *Nature* 559, 7714 (Gorffennaf 2018), 324-326.

Gweithredu Boniwr Porter ar gyfer y Gymraeg

VIGNESHWARAN MURALIDARAN

Ysgol Gwyddorau Cyfrifiadurol a Gwybodeg, Prifysgol Caerdydd, Cymru

GERAINT PALMER

Ysgol Fathemateg, Prifysgol Caerdydd, Cymru

LAURA ARMAN

Ysgol Saesneg, Cyfathrebu ac Athroniaeth, Prifysgol Caerdydd, Cymru

KEZIAH O'HARE

Ysgol Saesneg, Cyfathrebu ac Athroniaeth, Prifysgol Caerdydd, Cymru

DAWN KNIGHT

Ysgol Saesneg, Cyfathrebu ac Athroniaeth, Prifysgol Caerdydd, Cymru

IRENA SPASIĆ

Ysgol Gwyddorau Cyfrifiadurol a Gwybodeg, Prifysgol Caerdydd, Cymru

Geiriau ac ymadroddion allweddol: Prosesu iaith naturiol, Adfer gwybodaeth, Bonio, Lemateiddio

1 CYFLWYNIAD

Y broses o gwtogi gair i'w fôn yw bonio. Y ffurf graidd ar air yw'r bôn a gellir creu geiriau newydd ohono drwy atodi ychwanegiadau drwy brosesau ffurfdroi a deilliannau. Er enghraifft, byddai geiriau megis *gweithredu*, *gweithgar*, *gweithredoedd*, *gweithredol*, *gweithgaredd*, *gweithredwr* oll yn cael eu bonio i'r ffurf sylfaenol *gweith*. Yn yr un modd, byddai geiriau megis *symud*, *symudiad*, *symudwr*, *symudiadau*, *symudol* oll yn cael eu bonio i'r ffurf sylfaenol *sym*. Yn y rhan fwyaf o achosion (e.e. *gweith*) gyda rhai eithriadau (e.e. *sym*), bydd y bôn yn union yr un fath â gwraidd morffolegol y gair. Y gwreiddyn-air yw prif uned eiriol gair sy'n cario'r cynnwys semantig pwysicaf. Mae hefyd yn atomig yn yr ystyr na ellir ei ddadelfennu yn gydrannau llai. Felly, bydd bôn fel dirprwy i wreiddyn morffolegol yn adlewyrchu hanfod yr ystyr sy'n gysylltiedig â theulu cyfan o eiriau cysylltiedig (gweler yr esiamplau uchod). Dyma beth sy'n gwneud bonio yn werthfawr i gymwysiadau prosesu iaith naturiol (NLP) gan ei fod yn ei gwneud yn hawdd gwneud dadansoddiadau semantig heb orfod troi at adnoddau geiriol megis thesawrysau, nad ydynt bob amser ar gael yn hawdd, ac sy'n medru bod yn gymharol ddrud i'w datblygu.

Mae algorithmau bonio megis boniwr Porter [1] yn atebion ymarferol ysgafn i niwtraleiddio ffurfdroadau a deilliannau, gan ganolbwyntio felly ar yr ystyr greiddiol a gyflëir. Mae'r algorithm yn tynnu i ffwrdd ychwanegiadau (rhagddodiaid ac ôl-ddodiaid) yn ailadroddol gan roi sylw i'r newidiadau morffolegol a allai effeithio ar y ffurf sydd ar y wyneb. Er enghraifft, pan fo'r ychwanegiad *-ness* yn cael ei gyfuno â'r gair Saesneg *happy*, mae'r llafariad *y* yn newid yn *i*, h.y. *happy* + *-ness* = *happiness*. Felly, gan weithio at yn ôl, wrth dynnu'r ychwanegiad *-ness* o *happiness*, mae angen i'r llafariad *i* yn *happi* gael ei throï yn ôl yn *y*. Er mwyn delio â rhyngweithiadau o'r fath yn y rheolau bonio unigol mae gofyn astudio morffoleg yr iaith yn ofalus ac mae'r cam hwn yn gofyn am arbenigedd ieithyddol. Gellir datblygu rheolau i ddileu morffemau ffurfdroadol o air megis ffurfdroadau enwau neu rediadau berfau. Gyda dau air, W1 a W2, bydd yna wastad wahaniaeth barn ynglŷn ag a ddylid eu cydasio yn un bôn cyffredin ai peidio. Er enghraifft, gall y geiriau *perthnasol* a *perthynoledd* gael eu cydasio'n rhesymol i'r un bôn *perth*. Fodd bynnag, os yw'r corpws dan sylw yn cynnwys dogfennau sy'n ymwneud â ffiseg camgymeriad fyddai cydasio'r termau hyn yn un ffurf sylfaenol gan fod y gair *perthynoledd* ei hun yn gysyniad sydd ag ystyr arbennig na ddylid ei gydasio â geiriau eraill. Hyd yn oed os yw cyfradd llwyddiant boniwr yn llai na 100% oherwydd problemau o'r fath, mae datblygu algorithm syml a chyflym sy'n tynnu ymaith yr olddodiaid mwyaf cyffredin fel arfer yn gwella perfformiad y broses o adfer gwybodaeth.

2 GWAITH CYSYLLTIEDIG

Mae bonio gair yn cael ei ddefnyddio fel cam cyn prosesu mewn llawer o dechnegau adfer gwybodaeth yn enwedig algorithmau chwilio a mynegeo. Drwy gydasio nifer o amrywiadau ar air yn un bôn, gall algorithm chwilio drin holl amrywiadau'r gair fel elfennau perthnasol ar gyfer canlyniadau'r chwilio ac mae hyn yn gwella perfformiad yr algorithm chwilio. Mae lemateiddio yn dechneg debyg, sydd hefyd yn crynhoi gwahanol ffurfiau ar eiriau yn un ffurf graidd o'r enw lema. Mae lemateiddiwr yn adnabod y ffurf graidd drwy ystyried categori gramadegol y gair, yn aml drwy roi ystyriaeth i'r cyd-destun, ac yna mae'n esgor ar y lema ar sail y wybodaeth hon. Er enghraifft, byddai *gwneud*, *gwnaeth*, *gwnaed* oll yn cael eu cydasio i *gwneud* ar ôl lemateiddio gan eu bod yn perthyn i'r un rhan ymadrodd ac yn amrywiadau gramadegol ar yr un ferf. Ar y llaw arall, mae boniwr yn ceisio cydasio amrywiadau yn un ffurf graidd heb ystyried cyd-destun a chategorïau gramadegol. Un gwahaniaeth pwysig arall yw y dylai'r ffurf graidd a geir gan y lemateiddiwr fod yn ffurf eiriadurol neu'n ffurf forffolegol graidd ar air. Nid yw hyn yn wir am allbwn y boniwr. Felly, caiff *lleihau*, *lleihad*, *lleihawyd* eu cydasio yn *lleih* gan foniwr tra byddai lemateiddiwr yn adnabod *lleihau* fel y ffurf graidd.

Gellir grwpio dulliau bonio yn ddulliau talfyrru, dulliau ystadegol a dulliau cymysg [2]. Yn Saesneg ac mewn nifer o ieithoedd eraill, mae amrywiadau morffolegol yn digwydd ar ddiwedd ffurf ar air [3]. Mae algorithmau talfyrru yn cadw n o lythrennau mewn gair ac yn cael gwared ar y llythrennau ar ôl yr n-fed safle. Mae S-stemmer [4], boniwr Lovins [5] yn esiamplau o algorithmau talfyrru olddodiaid. Mae bonwyr Porters [1], Porters2 (neu boniwr Snowball) [6], [7], ac algorithm bonio Lancaster [8] yn algorithmau bonio blaenllaw sy'n cael gwared ar olddodiaid deilliannol a ffurfdroadol gair i gael y ffurf graidd. Ceir dulliau eraill o fonio sy'n gwneud dadansoddiadau ystadegol ar gorpws geiriau i adnabod y grŵp o eiriau y dylid eu cydasio i'r un ffurf wraidd. Mewn boniwr n-gram gwelir bod pob gair wedi'i wneud allan o gyfres o nodau. Drwy alldynnu n-gramau nodau ar gyfer y geiriau mewn corpws, mae'n bosibl cydasio amrywiadau ar yr un ffurf graidd oherwydd fe fydd gan eiriau tebyg nifer uchel o n-gramau yn gyffredin.

Cynigiodd Melucci ac Orio [9] ddull di-oruchwyliaeth o fonio gan ddefnyddio Model cudd Markov lle bo dilyniant llythrennau mewn gair yn cael eu modelu fel cyfres o set fôn a set olddodiaid. Caiff gair ei greu o ganlyniad i

drawsnewid rhwng y cyflyrau hyn. Er bod y dull hwn yn annibynnol ar iaith mae'n gymhleth ac weithiau yn gorfio'r geiriau. Ceir rhai dulliau, megis boniwr Krovetz [10], boniwr Xerox [11], sy'n gymysgedd o ddulliau seiliedig ar gorpws ochr yn ochr â dadansoddiadau o forffoleg ffurfdroadol a deilliannol. Cynigiodd Xu, J. a Croft, W.B. [12] ddull bonio seiliedig ar gorpws gan ddefnyddio cyd-ddigwyddiadau amrywiadau ar air. Mewn dulliau seiliedig ar gorpws, caiff y geiriau sydd â bôn cyffredin eu haddasu i gyd-fynd â nodweddion corpws testun penodol. Y syniad sylfaenol yw y dylai amrywiadau bôn gyd-ddigwydd yn y dogfennau o'r corpws. Cynigiodd Funchun Peng et al [13] foniwr sy'n sensitif i gyd-destun a fwriedir i'w ddefnyddio ar gyfer ymholiadau drwy chwilio'r we. Y syniad sylfaenol yw darogan amrywiadau defnyddiol ar air gan ddefnyddio boniwr arall megis boniwr Porter a chaiff yr amrywiadau hyn eu paru yn ôl cyd-destun er mwyn gwella ansawdd canlyniadau'r chwilio.

Mae i bob un o'r dulliau hyn eu manteision a'u hanfanteision eu hunain. Er nad oes yr un ohonynt yn bonio mewn ffordd gwbl berffaith, mae dewis y dull mwyaf priodol yn dibynnu ar yr iaith dan sylw, y cymhwysiad targed, a pha adnoddau ieithyddol sydd ar gael megis geiriadur a'r arbenigedd a ddarperir gan arbenigwr iaith. Allan o'r holl ddulliau a ddisgrifir uchod, algorithm gwreiddiol boniwr Porters, a gynigiwyd ar gyfer y Saesneg, yw'r boniwr a addaswyd yn ehangaf ar gyfer ieithoedd eraill. Gan mai dull ysgafn o fonio yw, nid yw'n gofyn am unrhyw strategaethau prosesu na strwythurau data cymhleth. Caiff y fethodoleg weithredu ei disgrifio yn adran 3.

3 METHODOLEG

Fe wnaethom weithredu algorithm bonio ar gyfer y Gymraeg drwy addasu egwyddorion algorithm bonio Porter [1], a weithredwyd yn wreiddiol ar gyfer y Saesneg, ond fe'i hail-weithredwyd ers hynny ar gyfer teulu mawr o ieithoedd Ewropeaidd [7]. Algorithm ailadroddol, seiliedig ar reolau ydyw, ac mae'n diffinio'n benodol set o olddodiaid, y meini prawf ar gyfer gweithredu'r rheolau dileu olddodiaid, a chwmpas a threfn gweithredu'r rheolau. Yr allwedd ar gyfer cynhyrchu algorithm bonio da ar gyfer iaith newydd yw creu sylfaen reolau briodol a'r drefn gywir ar gyfer gweithredu'r rheolau. Yn y project hwn, fe wnaeth tîm o ieithyddion (a oedd hefyd yn siaradwyr Cymraeg brodorol) ddatblygu rhestr wedi'i manwl gywreinio o reolau bonio (gweler Atodiad I am fanylion). O'i gymharu â'r Saesneg, lle gellir delio â ffurfdroadau drwy ganolbwyntio ar un gair ar y tro heb ystyried y geiriau cyfagos, nid yw'r dull di-gyd-destun hwn yn addas ar gyfer y Gymraeg, lle mae rhai newidiadau morffolegol (e.e. treigladau) yn sensitif i'r cyd-destun. I roi sylw i'r broblem hon yn effeithiol, fe wnaethom integreiddio lemateiddio i'r broses fonio. Yn benodol, rydym yn lemateiddio pob gair, h.y. yn ei fapio i'w ffurf ganonaidd (e.e. unigol ar gyfer enwau, berfenw ar gyfer berfau, ayb.), cyn gweithredu'r rheolau bonio. Mae hyn yn datrys y broblem "rhyngweithiad gair" ac yn caniatáu i bob gair gael ei fonio'n annibynnol ar ei gyd-destun. Yn ymarferol, mae'n lleihau nifer y rheolau bonio yn ogystal â chymhlethdod y ffyrdd o'u cyfuno, gan wneud yr algorithm bonio yn fwy cadarn ac yn haws ei gynnal. Caiff y lemateiddio yn ein dull ni ei gyflawni gan CyTag, sef tagiwr rhannau ymadrodd ar gyfer y Gymraeg a ddatblygwyd fel rhan o'r project CorCenCC [14].

Gellir crynhoi'r algorithm bonio Cymraeg gan ddefnyddio'r rhestr ganlynol o ddisgrifiadau o reolau lefel uchel a weithredir yn y drefn benodol:

1. Nodi cytseiniaid set C a berfau set V sy'n ffurfio sill yn y Gymraeg. Pan geir gair i'w fonio, dod o hyd i fesur geiriau 'm' a ddiffinnir fel sawl gwaith y mae'r dilyniant VC yn digwydd mewn gair.
2. Nodi set o reolau treiglo. Tynnu'r treigladau os yn berthnasol. Os yw canlyniadau amryfal yn bosibl, yna cyflawni proses ddadamwyso.
3. Dileu olddodiaid ffurfdroadol a deilliannol cenedl-rhif o air i esgor ar fôn enwau a berfau.
4. Dileu rhagddodiaid cymharol ac eithriadol i gael bôn ansoddeiriau.
5. Nodi rhestr o olddodiaid cynhyrchiol sy'n esgor ar enwau, berfau, ansoddeiriau ac adferfau o ffurf sylfaenol, e.e. *iau/au/od/ed/i/yn/ion/ydd/oedd*. Dileu'r olddodiaid hyn o air.

6. Ar gyfer camau 3-5 uchod, defnyddio'r mesur m i weld a yw'r rheol yn berthnasol ar gyfer gair neu beidio, e.e. ni chaiff y gair *tad* ei dalfyrnu i t drwy ddileu'r olddodiaid *-ad* gan nad yw ei fesur yn fwy nag 1.
7. Dileu rhagddodiaid cynhyrchol o'r gair.

Mae'r is-adrannau canlynol yn rhoi mwy o wybodaeth am y rheolau bonio.

3.1 Diffiniadau

Er mwyn disgrifio'r algorithm, mae angen y diffiniadau canlynol.

- **Cytsain C** – Caiff cytsain yn y Gymraeg ei diffinio'n ffurfiol fel unrhyw lythyren heblaw am a, e, i, o, u, y a'r llythyren w gyda'r amodau canlynol. Mae'r llythyren w yn gytsain pan nad yw wedi'i lleoli rhwng dwy gytsain neu pan gaiff ei dilyn gan lafariad. Er enghraifft, yn y gair *dwr*, mae'r llythyren w yn llafariad ond yn y gair *dwyr*, mae'r llythyren yn gytsain. Mae gan y Gymraeg wyth sain cytsain a ysgrifennir fel deugraffau (dwy lythyren) ond a ystyrir fel un llythyren. Dyma nhw: *ch, dd, ff, ng, ll, ph, rh, th*. Caiff y deugraffau eu hunain eu ffurfio o ddwy lythyren a chaiff y ddwy hon eu trin fel cytseiniaid yn unol â'n diffiniad. Caiff y deugraffau hyn eu trin fel un llythyren tra ein bod yn gweithredu rheolau bonio fel a ddisgrifir yn adran 3.2.
- **Llafariaid V** – Mae unrhyw beth nad yw'n gytsain yn llafariad.
- **Mesur m** – Gall unrhyw air gael ei gynrychioli ar ffurf $[C]VC\{m\}[V]$, lle bo C yn cynrychioli cyfres o gytseiniaid, mae V yn cynrychioli cyfres o lafariaid, mae $VC\{m\}$ yn cynrychioli cyfres o lafariaid a ddilynir gan gyfres o gytseiniaid wedi'u hailadrodd m gwaith. Yma, gelwir m yn fesur gair. Mae'r esiamplau canlynol yn dangos geiriau Cymraeg o wahanol fesurau m .
 - $m = 0$: y , $rhiw$, $trwy$, dau
 - $m = 1$: $gwynt$, bod , $llyfr$
 - $m = 2$: $gwyntoedd$, $byddaf$, $llyfrau$

Mae adnabod nifer y cyfresi VC mewn gair yn ddefnyddiol i gyflyru gweithrediad y rheolau dileu ychwanegiadau. Er enghraifft, os diffiniwn reol y gellir dileu'r ychwanegiad lluosog *au* o'r geiriau gyda $m \geq 2$, mae'n weithredol ar air fel *llyfrau* oherwydd bod ganddo fesur $m=2$. Mae hyn yn esgor ar fonion dilys *llyfr*. Fodd bynnag, ni fydd yr olddodiad *au* yn cael ei ddileu oddi ar *dau* oherwydd $m=0$. Gellir trefnu'r cyfryw reolau yn wahanol lefelau ar sail cwmmpas eu defnydd. Fel enghraifft, caiff rheolau talfyrnu olddodiaid eu gweithredu ar air cyn gweithredu'r rheolau dileu rhagddodiaid.

3.2 Rheolau Bonio ar gyfer Y Gymraeg

Caiff y diffiniadau o gytseiniaid a llafariad eu haddasu ar gyfer y Gymraeg i ddelio ag 'w' fel llafariad neu gytsain yn ddibynol ar y cyd-destun. Mae'r Gymraeg yn forffolegol gyfoethocach na'r Saesneg. At hynny, mae gan eiriau yn y Gymraeg reolau treiglo ar sail y geiriau sydd o'u blaenau, cenedl ramadegol y geiriau a'u swyddogaeth ramadegol. Er mwyn symleiddio'r dull o weithredu'r boniwr, bu i ni ddefnyddio lemateiddiwr sy'n ffurfio un gydran o CyTagger², sy'n dagiwr rhannau ymadrodd (POS) a grëwyd ar gyfer y Gymraeg. Defnyddiwyd y lemateiddiwr i fapio gair i'w ffurf ganonaidd cyn gweithredu'r rheolau bonio. Pryd bynnag yr oedd y lema ar gael o allbwn y lemateiddiwr, fe'i defnyddiwyd i symleiddio'r rheolau bonio. Gweithredwyd y rheolau bonio mewn dau gam: Mae Cam 1 yn cynnwys rheolau olddodiaid a weithredir yn gyntaf. Mae Cam 2 yn cynnwys rheolau rhagddodiaid a threigladau a weithredir nesaf. Mae cronfa'r boniwr Cymraeg ar gael ar Github³.

²<https://github.com/CorCenCC/CyTag>

³<https://github.com/CorCenCC/WelshStemmer>

Yng ngham 1, caiff yr olddodiaid cenedl-nifer eu dileu o'r gair i gael bonion enwau a berfau, caiff olddodiaid cymharol ac eithaf eu dileu i gael bonion ansoddeiriau a chaiff rhestr o olddodiaid cynhyrchiol sy'n esgor ar enwau, berfau, ansoddeiriau ac adferfau o ffurf sylfaenol eu dileu e.e. *iau/au/od/ed/i/yn/ion/ydd/oedd*. Yng ngham 2, gweithredir rhestr o reolau treigladau a rhagddodiaid cynhyrchiol. E.e. *cyd-, gwrth-, hunan-, rhag-, di-, an-, ad-, cyf-* ayb. Caiff rhestr fanwl o reolau rhagddodiaid, olddodiaid a threigladau eu disgrifio yn yr adrannau canlynol. Mae mesur *m* gair yn un ffordd yn unig o ganfod a yw rheol yn berthnasol mewn cyd-destun ai peidio. Dyma rai ffyrdd o bennu cwmpas gweithredu'r rheolau bonio:

- a yw'r gair yn dechrau â llafariad
- a yw'r gair yn dechrau â chytsain
- a oes llafariad yn y gair
- a yw'r gair wedi'i dreiglo ai peidio
- mesur *m* gair

Caiff y testun a fewnbynnir ei ragbroseu drwy CyTagger i gael y lemata, tagiau rhannau ymadrodd a niferoedd gramadegol y geiriau a fewnbynnir. Caiff y gair a fewnbynnir a'r wybodaeth arall eu prosesu drwy gamau 1 a 2 fel y disgrifir uchod. Mae'r gwahanol fathau o reolau a weithredir yn eu trefn i'w gweld yn yr is-adrannau canlynol. Yn yr is-adrannau mae SM yn cyfeirio at Dreigl Meddal, NM at Dreigl Trwynol ac mae AM yn golygu Treigl Llaes.

3.2.1 Dileu Ôl-ddodiaid

3.2.1.1 Ansoddeiriau sy'n deillio o enwau a berfau

Drwy ddileu'r olddodiaid o'r ansoddeiriau sy'n deillio o'r enwau a'r berfau, fe gawn fôn y gair neu'r ferf wreiddiol. Mae esiamplau ar gyfer y math hwn o reolau i'w gweld isod.

-(i)ol

anobeithiol	hopeless	(an + gobaith hope [SM])
beirniadol	critical	(beirniad ofnadwy)
cydwybodol	conscientious	(cydwybod conscience)
gogleddol	notherly	(gogledd north)
ieithyddol	linguistic	(ieithydd linguist)

-aidd

deuaidd	binary	(dau two)
oeraidd	wintry	(oer winter)
niwlaidd	fuzzy, blurred	(niwl haze, fog)

Mae enw a roddir ynghlwm wrth enw arall mewn gair cyfansawdd yn achosi treigl meddal (SM). Rhaid delio â hyn wrth ganfod y bôn.

llys (court) + mam (mother) = llysfam

3.2.1.2 Olddodiaid ffurfdroi nifer-cenedl ar fôn enw

Caiff yr olddodiaid sy'n dynodi cenedl a nifer eu dileu i gael bôn yr enw. Y rhai mwyaf cyffredin yw: *IAU/AU/OD/ED/I/YN/ION/YDD/OEDD*.

Ni chaiff yr olldodiad -ACH ei ddileu o'r ansoddair gan ei fod eisoes yn fôn.

EITHRIADAU: rhagorach -> rhagor
 cyfrinach -> cyfrin

3.2.1.3 Ansoddair y radd eithaf + -ach

Caiff olldodiaid sy'n nodi graddau cymharol ac eithaf ansoddair eu dileu i gael ffurf graidd yr ansoddair.

E.E. MWYACH -> MWY
EITHRIADAU: HYTRACH, CHWAETHACH

Os yw'r gair sy'n gorffen mewn -ACH yn ENW, nid oes angen ei ddileu. Mae eisoes yn fôn. Gwelwyd yr eithriadau canlynol ar gyfer y rheol hon.

EITHRIADAU:
CLOGYRNACH -> CLOGYRN
CYFEDDACH -> CYFEDD
CYFEILLACH -> CYFEILL
CHWANTACH -> CHWANT
CREPACH -> CREP
CRWBACH -> CRWB
CRWMAC -> CRWM
GWYACH -> GWY
LLINACH -> LLIN
LLOSGACH -> LLOSG
SIMACH -> SIM
SINACH -> SIN
SOTHACH -> SOTH
SWBACH -> SWB
TOLACH -> TOL

Fel arall, gellir dileu'r olldodiad.

-ACH PETHEUACH -> PETHEU
 YSBLEDDACH -> YSBLEDD
 SIAFFLACH -> SIAFFL
 PLANTACH -> PLANT
 MERCHETACH -> MERCHET

-IACH BLEWIACH -> BLEW
 HOLLIACH -> HOLL

EITHRIADAU:
AFIACH -> AFIACH

-AETH	DILEWCH AR GYFER ENWAU A BERFENWAU YN UNIG
MEDDYGAETH	-> MEDDYG
ALAETH	-> ALAETH

PWYSIG -AETH/-IAETH

Mae'n bwysig tynnu'r olddodiad -**IAETH** cyn dileu'r ôl-ddodiad -**AETH**. Mae trefn gweithredu'r rheolau yn sicrhau y bwrir cyfrif o'r olddodiad -**IAETH** i ddechrau a'i fod yn cael ei dynnu cyn dileu -**AETH**.

-IAETH	Caiff hwn ei dynnu ar gyfer enwau yn unig
E.E. CYFRIFIAETH	-> CYFRIF
MILWRIAETH	-> MILWR

3.2.1.4 Olddodiad eraill

-AS	PRIODAS	-> PRIOD
-EB	TYSTEB	-> TYST
-EG	CYMRAEG	-> CYMRA

-ELL	Ôl-ddodiad bachigol	
BRIWELL	-> BRIW	CRAFELL-> CRAF
BOTYMELL	-> BOTYM	BYSELL -> BYS
TERFYNELL	-> TERFYN	

Mae tynnu'r olddodiad lluosog -**AU** yn arwain at fôn unigol. Caiff hyn wastad ei ddefnyddio ar ansoddeiriau neu enwau lluosog.

E.G. AMLINELLAU -> AMLINELL

Ar ferfau ac ansoddeiriau unigol, ni thynnir yr ôl-ddodiad hwn i ffwrdd. Ceir rhai eithriadau i hyn.

EITHRIADAU:

AMLHAU -> AML, AGOSÁU -> AGOS, ANNETHAU -> ANNETH, DADLAU -> DADL (GALL Y GAIR HWN FOD YN ENW HEFYD, YN DDIBYNNOL AR Y CYD-DESTUN), DETHAU -> DETH

Fodd bynnag, ar enwau unigol, caiff yr ôl-ddodiad ei dynnu i ffwrdd.

-EN	SEREN -> SER, COEDEN -> COED, GWENYNEN -> GWENYN (NIFER)
-ES	YSGRIFENYDDDES -> YSGRIFENYDD, ARGLWYDDDES -> ARGLWYDD, BRENHINES -> BRENIN (CENEDL)
-ES	BUCHES -> BUCH
-FA	GRADDFA -> GRADD
-IN	CRIBIN -> CRIB
-OEDD	DANNOEDD -> DANN
-OR	YSGUBOR -> YSGUB

-RED	GWEITHRED	-> GWEITH
-WRAIG	GWEITHWRAIG	-> GWEITH

OLDDODIAID GWRYWAIDD

-AD	CARIAD	-> CARI
-ADUR	HOLIADUR	-> HOLI
-AI	MABWYSIADAI	-> MABWYSIAD
-AINT	HENAINT	-> HEN
-AWD	PELAWD	-> PEL
-AWDWR	CREAWDWR	-> CRE
-CYN	BRYNCYN	-> BRYN
-DEB	CYTUNDEB	-> CYTUN
-DER	CYFIAWNDER	-> CYFIAWN
-DID	CALONDID	-> CALON
-DOD	UFUDD-DOD	-> UFUDD
-DRA	BUANDRA	-> BUAN
-DWR	ADEILADWR	-> ADEILAD
-EDD	BRWDFRYDEDD	-> BRWDFYD
-FEL	OERFEL	-> OER
-I	CALEDI	-> CALED
-IANT	FFYNIANT	-> FFYN
-ID	RHYDDID	-> RHYDD
-INEB	GWARINEB	-> GWARIN
-MON	POSTMON	-> POST
-OD	RHYFEDDOD	-> RHYFEDD
-OL	GOLYGYDDOL	-> GOLYGYDD
-OR	PORTHOR	-> PORTH
-RWYDD	AEDDFEDRWYDD	-> AEDDFED
-WCH	AFLERWCH	-> AFLER
-WR	GARDDWR	-> GARDD
-YCH	EURYCH	-> EUR
-YD	CLEFYD, IECHYD	-> CLEF, -> IECH
-YDD	CYNORTHWYDD	-> CYNORTHWY
-YN	ADERYN	-> ADER

Y NAILL NA'R LLALL

-AID	CWPANAID	-> CWPAN
------	----------	----------

ONI BAI EI FOD YN -RHAI, RHODDIR -RAID YNGHLWM WRTH UN SILL (GAN MAI GEIRIAU CYFANSAWDD YW'R RHAIN).

-AN	BLEIDDIAN, GWREIGAN	-> BLEIDD, -> GWREIG
-----	---------------------	----------------------

-ED	COLLED,SYCHED	-> COLL,-> SYCH
-IG	MORWYNIG ,PENDEFIG	-> MORWYN,-> PENDEF
-OD	DYRNOD,FFEDOG-	> DYRN,-> FFED
-OG	CYMYDOG	-> CYMYD,
-YLL	GWYNTYLL	-> GWYNT

ENWAU LLUOSOG

-OS PLANTOS	-> PLANT
-------------	----------

LLUOSOG AC UNIGOL

-ACH	DYNIONACH	-> DYNION (LL.)
-ACH	CORRACH	-> CORR
-ACH	CLINDDARACH CYFFEILLACH	-> CLINDDAR, -> CYFFEILL (BERFENW)

Ar gyfer geiriau sy'n gorffen **-NT** pan maent angen **-NN** yn y lluosog, cafodd popeth ar ôl **-NN** ei dynnu i ffwrdd ac ychwanegwyd **-NT**.

CANNOEDD	-> CA	-> CANT
DANNEDD	-> DA	-> DANT
DIWYLLIANNAU	-> DIWYLLIA	-> DIWYLLIANT

Eithriadau: GWYNT, PONT, RHIANT

3.2.2 Cam 2 – Tynnu Rhagddodiaid

3.2.2.1 Enwau a berfau

Caiff y rhagddodiaid deilliannol canlynol eu dileu o enwau a berfau. Dylid nodi y gellir sillafu'r rhagddodiaid hyn gyda chysylltnod neu beidio.

m > 0, cyd

cyd- co-,con-:

CYDBWYSEDD	BALANCE	PWYSEDD WEIGHT, PRESSURE
CYD-DDIGWYDDIAD	COINCIDENCE	DIGWYDD HAPPEN
CYDWEITHWYR	COLLEAGUES	GWEITHWYR WORKERS
CYDFYW	COHABIT	BYW LIVE
CYDFYND	ACCOMPANY	MYND GO

Cydym-

gwrth- anti-, counter-, against:

GWRTHBLAID	OPPOSITION PARTY	PLAID (GWLEIDYDDOL) PARTY
GWRTHGYNHRYCHOL	COUNTERPRODUCTIVE	CYNHYRCHOL PRODUCTIVE

hunan- self

HUNANBARCH	SELF-ESTEEM	PARCH RESPECT
------------	-------------	---------------

HUNANLADDIAD	SUICIDE	LLADD KILL
rhag- pre-, fore-		
RHAGWELD	FORESEE	GWELD SEE
RHAGFARN	PREJUDICE	BARN JUDGMENT
ym - [Sy'n aml yn golygu <i>hunan</i> neu <i>ei gilydd</i>]		
YMOLCHI	WASH (ONSELF)	GOLCHI WASH
YMLADD	FIGHT	LLADD KILL

3.2.2.2 *Ansoddeiriau*

AF- UN-:

AFRESYMOL	UNREASONABLE	RHESYMOL REASONABLE
AFLWYDDIANNUS	UNSUCCESSFUL	LLWYDDIANNUS SUCCESSFUL

DI- UN-, -LESS, WITHOUT:

DIDRAFFERTH	WITHOUT PROBLEMS	TRAFFERTH TROUBLE, PROBLEMS
DI-GYMRAEG	NON-WELSH SPEAKING	WELSHLESS
DIBAID	CEASELESS	PAID, PEIDIO CEASE

EITHRIAD: DIABETIG, DIACRONIG, DIACEN, DIAFAEL

- RHAGDDODIAID FFURFDROADOL SY'N ACHOSI TREIGLAD MEDDAL (SM)

AF- UN-:

AFRESYMOL	UNREASONABLE	RHESYMOL REASONABLE
AFLWYDDIANNUS	UNSUCCESSFUL	LLWYDDIANNUS SUCCESSFUL

DI- UN-, -LESS, WITHOUT:

DIDRAFFERTH	WITHOUT PROBLEMS	TRAFFERTH TROUBLE PROBLEMS
DI-GYMRAEG	NON-WELSH SPEAKING	WELSHLESS
DIBAID	CEASELESS	PAID, PEIDIO CEASE

EITHRIAD: DIABETIG, DIACRONIG, DIACEN, DIAFAEL

- RHAGDDODIAID DEILLIANNOL SY'N ACHOSI TREIGLAD TRWYNOL (NM)

AN-

ANGHOFIO	FORGET + COFIO	REMEMBER
----------	----------------	----------

- RHAGDDODIAID FFURFDROADOL SY'N ACHOSI TREIGLAD TRWYNOL MEWNOL (NM)

AN-

ANNHEBYG	UNLIKELY	+ TEBYG	LIKELY
ANNARLLENADWY	ILLEGIBLE	+ DARLLENADWY	LEGIBLE
ANGHYSON	INCONSISTENT	+ CYSON	CONSISTENT
AMHOSIB	IMPOSSIBLE	+ POSIB	POSSIBLE

Caiff geiriau treigladrwynol eu hystyried yn rhai cysefin. Nid oes ond un achos mewn Cymraeg Modern pan gaiff y treigladrwynol ei weithredu'n gyson.

SYLWER:

-Mae'r N o AN- yn diflannu wrth dreiglo B- i M-, C- i NGH- neu P- i MH-:

AMHENDANT (INDEFINITE)

AN + PENDANT (DEFINITE)

Mae pob gair cysefin sy'n dechrau â TR- yn achosi i un o'r N'au deilliannol ddiflannu.

ANHREFN (CHAOS) NID *ANNHREFN

AN + TREFN (ORDER)

Gellir gweld ym mha drefn y caiff ôl-ddodiaid a rhagddodiaid eu tynnu o'r bôn o'r cod bonio sydd ar gael ar-lein. Ar ôl dileu'r ôl-ddodiaid a'r rhagddodiaid, caiff y rheolau treiglo eu gweithredu i gael y ffurf graidd, wreiddiol heb ei threiglo.

4 GWERTHUSO

Caiff effeithlonrwydd bonio yn aml ei fesur yn ôl y gostyngiad bach ym maint mynegai a geir drwy fonio. Caiff y ffactor cywasgu mynegai (ICF) ei gyfrifo yn unol â'r fformiwla ganlynol:

$$ICF = (W - S) / W$$

Lle bo w yn golygu nifer y geiriau unigol cyn bonio ac s yw nifer y bonion unigol [15].

Bu i ni defnyddio CorCenCC [16], sy'n gorpws cenedlaethol Cymraeg cyfoes, a'i seilwaith cyfrifiadurol [17] i gyflawni ein harbrofion. Rhagbroseswyd corpws CorCenCC i beidio â chynnwys atalnodau, symbolau, rhifau, tagiau di-enw, termau tramor, cyfeiriadau gwe, rhifau ffôn, emojis, a chyfeiriadau e-bost. Ar ôl y cam hwn, ceid cyfanswm o 10,276, 495 o docynnau yn y corpws. Roedd nifer y geiriau gwahanol yn y corpws yn $W = 105,780$ gyda nifer y lemata gwahanol $L = 81,347$. Ar ôl bonio'r lemata, roedd nifer y bonion gwahanol yn $S = 70,153$. Cafodd yr ICF ar ôl lemateiddio a bonio yn y drefn honno ei gyfrifo fel a ganlyn:

$$ICF1 = (W - L) / W = (105,780 - 81,347) / 105,780 = 0.231$$

$$ICF2 = (W - S) / W = (105,780 - 70,153) / 105,780 = 0.337$$

Y gostyngiad ffracsiynol ym maint y mynegai rhwng lemateiddio a bonio yw:

$$ICF3 = (L - S) / L = (81,347 - 70,153) / 81,347 = 0.138$$

Er mwyn gwerthuso'r boniwr yn ansoddol, mae Tabl 1 yn rhoi rhestr o'r bonion a gynhyrchwyd o'r geiriau a welir amlaf yn y corpws CorCenCC. Mae'r golofn gyntaf yn darparu'r 10 gair sy'n ymddangos amlaf nad ydynt yn newid dan y rheolau bonio, mae'r ail golofn yn rhoi'r 10 gair sy'n digwydd amlaf sy'n bonio i eiriau sydd eisoes i'w cael yn y corpws, ac mae'r drydedd golofn yn rhoi'r 10 gair mwyaf cyffredin sy'n bonio i ffurfiau ar eiriau nad oeddent i'w cael yn y corpws yn flaenorol. Mae sgôr amllder cyffredinol y bonion hefyd yn cael ei rhoi mewn cromfachau. Mewn ieithoedd sydd â morffoleg gyfoethog iawn, byddai rhywun yn disgwyl i gyfran fawr o'r bonion hefyd fod yn eiriau ynddynt eu hunain.

Tabl 1 Gwerthusiad ansoddol o allbwn y boniwr

Dim Newid	I Eiriau	I Newydd
yn - yn (1)	'r - y (5)	cymru - cymr (32)
i - i (2)	'n - yn (7)	o'n - o yn (69)

y - y (3)	mae - bod (9)	gymraeg - gymra (70)
a - a (4)	yr - y (10)	i'w - i ei (73)
o - o (6)	ac - a (12)	cymraeg - cymra (75)
ar - ar (8)	oedd - bod (19)	dros - tros (80)
ei - ei (11)	sy - bod (26)	wedyn - wed (92)
wedi - wedi (13)	yw - bod (28)	achos - ach (112)
am - am (14)	'i - ei (29)	siarad - siar (124)
ni - ni (15)	fod - bod (30)	amser - ser (126)

Dengys y tabl bod y geiriau sy'n digwydd amlaf naill ai ddim yn newid neu'n bonio yn eiriau hysbys, tra bo dim ond 7 o'r 100 gair uchaf yn cynhyrchu geiriau nas gwelwyd. Yma, mae'r holl eiriau nad ydynt yn newid, y golofn gyntaf, yn cyfateb i eiriau nad oes arnynt angen newid gan eu bod ar eu ffurf symlaf. Mae'r ail golofn yn tynnu sylw at y ffaith bod gan rai o'r geiriau sy'n ymddangos amlaf forffoleg afreolaidd, gyda'r rhan fwyaf yn ffurfiau amrywiol ar *bod*. Cafodd y geiriau hyn yn wir eu bonio yn gywir ac yn synhwyrol. Dengys y golofn olaf y geiriau sy'n cynhyrchu bonion nad ydynt yn eiriau ynddynt eu hunain. Mae llawer o'r rhain yn gwneud synnwyr, fodd bynnag, mewn rhai achosion, er enghraifft, *achos* – *ach*, *siarad* – *siâr* ac *amser* – *ser*, nid yw'r boniwr wedi cynhyrchu'r bôn amlycaf. Anaml yw'r holl achosion hyn, fodd bynnag, ac nid oes yr un ohonynt yn ymddangos yn y 100 gair uchaf yn ein corpws. Hefyd, mae'n amlwg sut mae'r boniwr wedi cam-fonio'r rhain. Maent oll yn cynnwys olddodiaid neu ragddodiaid cyffredin (*os*, *ad*, ac *am*), a fyddai yn y rhan fwyaf o achosion eraill yn ddewisol i'w tynnu o'r gair.

5 CASGLIADAU

Mae'r bennod hon yn disgrifio sut aed ati i addasu boniwr Porter ar gyfer y Gymraeg. Mae bonio yn un o offer sylfaenol rhagbrosesu ieithyddol ac mae'n un o'r prif alluogwyr ar gyfer datblygu technoleg iaith dynol mewn iaith nad oes ganddi adnoddau digonol fel y Gymraeg. Mae'r boniwr yn cyd-fynd â'r gweithgareddau prosesu iaith naturiol sy'n mynd rhagddynt eisoes gyda'r Gymraeg. Ar hyn o bryd, nid oes yr un o'r offer presennol yn cefnogi bonio. Er enghraifft, mae Cit Offer Iaith Naturiol y Gymraeg⁴ yn gweithredu set o offer prosesu iaith naturiol y Gymraeg ar gyfer tocneiddio, lemateiddio a thagio rhannau ymadrodd. Cafodd set o offer gyda galluoedd NLP tebyg eu gweithredu i helpu i rhagbrosesu dogfennau a storiwyd yn CorCenCC, sydd wedi'u tocneiddio gan ddefnyddio CyTag, sy'n ddiwr rhannau ymadrodd seiliedig ar reolau, sydd hefyd yn lemateiddio [14]. Mae tagiwr rhannau ymadrodd arall, y gellir ei ddefnyddio fel gwasanaeth rhyngwyd heb fod angen ei lwytho'n lleol, yn gallu tagio categorïau geiriol (e.e. berfau ac enwau), ond hefyd yn benodol nodweddion y Gymraeg fel treigladau [18]. Fe wnaeth yr un tîm ddatblygu lemateiddiwr, y gellir ei ddefnyddio i normaleiddio unrhyw air sydd wedi'i ffurfdroi, ei dreiglo a/neu ei redeg i'w lema [19]. Gellir cyfuno ein boniwr Cymraeg ag unrhyw rai o'r offerynnau hyn yn ogystal â mewnbaniadau geiriau Cymraeg [20] i wella agweddau semantig ar brosesu iaith naturiol yn Gymraeg.

⁴<https://sourceforge.net/projects/wnlt-project/>

DIOLCHIADAU

Ariannwyd yr astudiaeth hon gan Lywodraeth Cymru fel rhan o broiect a oedd yn ymwneud â seilwaith prosesu'r iaith Gymraeg.

CYFEIRIADAU

- [1] M. F. Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130-137.
- [2] Anjali G. Jivani. 2011. A comparative study of stemming algorithms. *International Journal of Computer Applications in Technology* 2, 6 (2011), 1930-1938.
- [3] Richard W. Sproat a Maurice V. Wilkes. 1992. *Morphology and computation*. MIT press.
- [4] Donna Harman. 1991. How effective is suffixing?. *Journal of the american society for information science* 42, 1 (Ionawr 1991), 7-15.
- [5] Julie B. Lovins. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* 11, 1-2 (Mawrth a Mehefin 1968), 22-31.
- [6] M.F.Porter. 2001. Snowball: A language for stemming algorithms. Cyrchwyd Mai 16, 2021 o <http://snowball.tartarus.org/texts/introduction.html>
- [7] M. F. Porter. Stemming algorithms for various European languages. Cyrchwyd Mai 16, 2021 o <http://snowball.tartarus.org/texts/stemmersoverview.html>
- [8] C. Paice, A R. Hooper. 2005. Lancaster stemmer. <https://github.com/words/lancaster-stemmer>
- [9] Massimo Melucci a Nicola Orio. 2003. A novel method for stemmer generation based on hidden Markov models. Yn Proceedings of the twelfth international conference on Information and knowledge management. Association for Computing Machinery, Efrog Newydd, NY, UDA, 131-138.
- [10] Robert Krovetz. 2000. Viewing morphology as an inference process. *Artificial intelligence* 118, 1-2 (Ebrill 2000), 277-294.
- [11] David A. Hull a Gregory Grefenstette. 1996. A detailed analysis of English stemming algorithms. Technical Report. Rank Xerox Research Centre.
- [12] Jinxi Xu a W. Bruce Croft. 1998. Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems*, 16, 1 (Ionawr 1998), 61-81.
- [13] Fuchun Peng, Nawaaz Ahmed, Xin Li a Yumao Lu. 2007. Context sensitive stemming for web search. Yn Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. Association for Computing Machinery, Efrog Newydd, NY, UDA, 639-646.
- [14] Steve Neale, Kevin Donnelly, Gareth Watkins a Dawn Knight. 2018. Leveraging Lexical Resources and Constraint Grammar for Rule-Based Part-of-Speech Tagging in Welsh. Yn Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). LREC, Miyazaki, Japan.
- [15] William B. Frakes a Christopher J.Fox. 2003. Strength and similarity of affix removal stemming algorithms. Yn ACM SIGIR Forum (Vol. 37, No. 1). ACM, Efrog Newydd, NY, UDA, 26-30.
- [16] Dawn Knight, Steve Morris, Tess Fitzpatrick, Paul Rayson, Irena Spasić ac Enlli-Mon Thomas. 2020. The National Corpus of Contemporary Welsh: Project Report | Y Corpws Cenedlaethol Cymraeg Cyfoes: Adroddiad y Prosiect. Prifysgol Caerdydd<18/>, Cymru.<19/>. arXiv:2010.05542.
- [17] Dawn Knight, Fernando Loizides, Steven Neale, Laurence Anthony ac Irena Spasić. 2020. Developing computational infrastructure for the CorCenCC corpus: The National Corpus of Contemporary Welsh. *Language Resources and Evaluation*. <https://doi.org/10.1007/s10579-020-09501-9>
- [18] Dewi Bryn Jones, Patrick Robertson a Gruffudd Prys. 2015. Welsh language Parts-of-Speech Tagger API Service. Uned Technolegau Iaith Canolfan Bedwyr, Prifysgol Bangor, Bangor, Cymru. <http://techiaith.cymru/api/parts-of-speech-tagger-api/?lang=en>
- [19] Dewi Bryn Jones, Patrick Robertson a Gruffudd Prys. 2015. Welsh language lemmatizer API service. Uned Technolegau Iaith Canolfan Bedwyr, Prifysgol Bangor, Bangor, Cymru. <http://techiaith.cymru/api/lemmatizer/?lang=en>
- [20] Geraint Palmer, Padraig Corcoran, Laura Arman, Dawn Knight ac Irena Spasic. 2021. A closer look at Welsh word embeddings. Iaith a Thechnoleg yng Nghymru: Cyfrol 1. Uned Technolegau Iaith Canolfan Bedwyr, Prifysgol Bangor, Bangor, Cymru.

Datblygu ColloCaid: offer delweddu a golygu testun i helpu ysgrifenywyr â chydleoliadau

JONATHAN C. ROBERTS

Ysgol Gwyddorau Cyfrifiadurol a Pheirianeg Electronig, Prifysgol Bangor, Cymru

PETER W. S. BUTCHER

Ysgol Gwyddorau Cyfrifiadurol a Pheirianeg Electronig, Prifysgol Bangor, Cymru

ROBERT LEW

Cyfadrn Saesneg, Prifysgol Adam Mickiewicz, Poznań, Gwlad Pwyl

GERAINT REES

Canolfan Astudiaethau Cyfieithu, Ysgol Llenyddiaeth ac Ieithoedd, Prifysgol Surrey, Lloegr

NIRWAN SHARMA

Cyfadrn Gwyddoniaeth, Technoleg, Peirianeg a Mathemateg, Prifysgol Agored, Lloegr

ANA FRANKENBERG-GARCIA

Canolfan Astudiaethau Cyfieithu, Ysgol Llenyddiaeth ac Ieithoedd, Prifysgol Surrey, Lloegr

Rydym wedi datblygu'r golygydd testun ColloCaid, sy'n offer cynorthwyo ysgrifennu sy'n awgrymu cydleoliadau ar gyfer geiriau wrth i'r defnyddwyr eu teipio. Mae'r offer hefyd yn delweddu geiriau i egluro cysylltiadau geiriau ac i gyflwyno geiriau amgen. Yn ystod y gwaith o ddylunio, datblygu, a gwerthuso offer delweddu a golygydd bu'n rhaid delio â nifer o ddewisiadau, heriau a phenderfyniadau. Drwy fod ColloCaid yn broiect cydweithredol, fe wnaeth ddwyn ynghyd ymchwilwyr a oedd yn meddu ar wahanol sgiliau, cefndiroedd ac arddulliau gweithio. Yn y bennod hon, rydym yn trafod ein cylch datblygu wrth greu'r offer ColloCaid, o ddewis data corpws y testun cyfeirio i ddylunio a datblygu delweddiadau geiriau priodol. Gall ein hadroddiad fod o ddiddordeb i ymchwilwyr eraill, gan gynnig cipolwg ar ddatblygu offer ieithyddol rhyngweithiol sy'n cynnwys delweddiadau.

Ymadroddion a geiriau allweddol: delweddu, ieithyddiaeth, ieithyddiaeth corpws, ysgrifennu agos

1 CYFLWYNIAD

Wrth ysgrifennu testunau academiaidd, gall dewis y geiriau iawn fod yn dipyn o her. Yn y gwaith hwn rydym yn canolbwyntio ar helpu ysgrifenywyr i ddewis y geiriau gorau, yn benodol, yn hwyluso mynediad at bartneriaid geiriol arferol: cydleoliadau cryf (*strong collocations*). Gyda'r pwrpas hwn mewn cof, rydym wedi datblygu'r offer delweddu geiriau a'r golygydd testun gan ddefnyddio cronfa ddata eiriol a grëwyd o gorpora testun o safon i gynnig i ysgrifenywyr ddetholiad o gydleoliadau wrth iddynt weithio ar eu testun.

Gall delio â thestun newydd fod yn anodd i ysgrifenywyr. Mae awduron a fu'n academyddion ers tro yn tueddu i fod yn gyfarwydd â'r geiriau a'r ymadroddion a ddefnyddir yn gyffredin yn y byd academiaidd. Mae eu testun ysgrifenedig yn llifo'n dda, mae iddo strwythur sy'n briodol i'r maes ac mae'r geiriau allweddol a ddefnyddiant yn berthnasol i'r testun. Ar y llaw arall, gall ysgrifennu academiaidd fod yn ddychrynlyd i ysgrifenywyr newydd, felly hefyd ysgrifenywyr nad ydynt yn ysgrifennu yn eu hiaith gyntaf, gan nad ydynt o anghenraid yn gwybod y geiriau neu'r ymadroddion sy'n arferol yn y maes hwnnw. Fe allent o bosibl ddewis defnyddio gair sydd ag ystyr tebyg, ond

nas defnyddir yn gonfensiynol yn y cyd-destun hwnnw, neu ddefnyddio geiriau nad ydynt yn mynd yn dda gyda'i gilydd fel arfer, gan gynhyrchu cydleoiliad geiriol amhriodol. Mae Krishnamurthy [1] yn diffinio cydleoiliad fel geiriau sy'n digwydd "with a greater frequency than the law of averages would lead you to expect". Mae'r ymadroddion yn ymddangos oherwydd y cânt eu defnyddio drosodd a drosodd mewn cyd-destun penodol. Mae gwybodaeth am gydleoiliadau yn bwysig er mwyn cyflwyno brawddegau naturiol (ac felly clir). Mae dewis cydleoiliadau anghywir yn arwain at ryddiaith drwsgl, neu fe all ymddangos yn ddieithr, gan dynnu sylw darllenwyr oddi ar y cynnwys.

Mae cydleoiliadau yn digwydd ym mhob maes, er nad yr un cydleoiliadau ydynt o anghenraid. Er enghraifft, byddai rhywun sy'n ysgrifennu am gyfrifiaduron yn dweud bod ganddynt *gyfrifiadur pwerus* o bosibl – cyfrifiadur sydd â sawl petafflop o bŵer cyfrifiadurol a nifer o greiddiau cyfrifiadura. Er y byddai modd i ansoddeiriau eraill gael eu defnyddio yn gyffredinol yn lle *pwerus*, megis *cyhyrol*, *cryf*, neu *gadarn*, fel y gwelir mewn thesawrws, byddent naill ai'n amhriodol neu'n cyfleu ystyr gwahanol. Felly, byddai **cyfrifiadur cyhyrol* yn gyfuniad geiriol amwys ac anghyffredin iawn, tra byddai *cyfrifiadur cadarn* yn golygu rhywbeth gwahanol: cyfrifiadur sy'n gallu gwrthsefyll straen ffisegol, er nad yw o anghenraid yn un sy'n perfformio'n gyflym. Ym maes delweddu, mae awduron yn defnyddio termau megis *siart bar* i gyfeirio at *plot bar*. Pam felly? Yn bendant, fe ellid cyfnewid y geiriau *siart* a *plot*, ond mae'n well gan awduron *siart cylch* na **plot cylch* [2]. Maent ganddynt ystyron tebyg iawn, a gallai'r ddwy fersiwn fod yn briodol. Ond mae awduron eraill yn y maes yn nodweddiadol wedi dewis un ffordd, yn hytrach nag un arall, i fynegi eu syniadau. Po amlaf y deuen ar draws cyfuniad geiriau, po fwyaf cyfarwydd, sefydledig, ac felly naturiol ydyw, ac mae'n tyfu'n gydleoiliad cryf. Mae cydleoiliadau gwan yn gwbl groes: maent yn llai cyffredin, yn swinio'n annaturiol ac yn gwneud y testun yn chwithig i'w ddarllen.

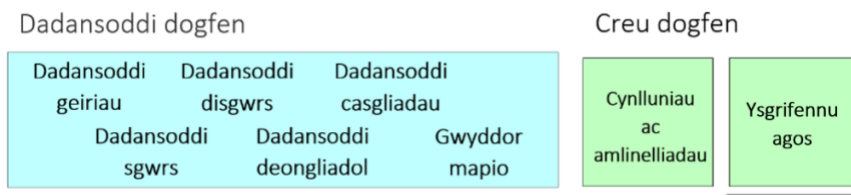
Gellir dadansoddi cydleoiliadau drwy archwilio casgliadau digidol mawr o destunau sy'n bodoli eisoes sy'n gweithio fel tystiolaeth o ddefnydd iaith. Mae'r broses ddadansoddi yn llwytho testunau y gall peiriant eu darllen a, gan ddefnyddio modelau ystadegol, yn darparu i'r defnyddiwr wybodaeth ddefnyddiol am y gair. Yn y sefyllfa hon, mae'r data a fewnbynnir (y testunau y gall peiriant ei ddarllen) a'r data a allbynnir (nodweddion ystadegol a dosbarthiadol y geiriau) yn helpu'r ysgrifennwr i ddeall y maes. O'r herwydd, gall ysgrifennwr chwilio gair i ddod o hyd i'w gydleoiliad confensiynol, yn ogystal â defnyddio'r testunau gwreiddiol i archwilio brawddegau enghreifftiol i'w helpu i ddewis y termau priodol: "corpus linguistics is not concerned with what is possible in a language, but in [sic] what is probable" [3]. Diolch i gorpora, gall ysgrifennwyr ddewis y geiriau y mae arnynt eu hangen i fynegi eu syniadau'n effeithiol [3], [4]. Mae meddalwedd chwilio corpws yn tueddu i fod yn ddigon pwerus i gyflwyno gwybodaeth gynhwysfawr am y corpws dadansoddedig, ond eto mae'n rhy anodd i'w ddefnyddio i rai nad ydynt yn ieithyddion. Ni fydd y rhan fwyaf o ddefnyddwyr yn ymwybodol o sut gall y corpora fod yn ddefnyddiol iddynt, na sut i greu ymholiadau perthnasol i'w helpu i ysgrifennu. Mae hefyd yn hawdd mynd â sylw ysgrifennwyr a gallent ddilyn geiriau amherthnasol, ac fe allant o bosibl gamddehongli ystyr y canlyniadau. Yr hyn sy'n ofynnol yw ffordd o integreiddio manteision offer ieithyddol corpws gyda golygydd.

Drwy ddatblygu golygydd testun penodol sy'n cynnwys dadansoddi'r iaith a'r corpws, byddai ysgrifennwr yn archwilio cydleoiliadau penodol ac yn gweld digon o ddewisiadau amgen i wneud dewis gwybodus, ond eto nid gormod rhag peri dryswch. Gall delweddu priodol hefyd helpu i gyflwyno'r dewis o eiriau amgen yn effeithiol. Gyda'r offeryn hwn, gallai defnyddwyr archwilio'r cydleoiliadau *in-situ* a pheidio â symud i ffwrdd o'r golygydd oherwydd fe allai hynny atal llif eu hysgrifennu; a chael mynediad at ddulliau sy'n eu hannog i ymchwilio ac archwilio geiriau newydd a gwella'u sgiliau ysgrifennu. Yn y bennod hon, cyflwynwn yr offeryn ColloCaid gan egluro sut cafodd ei ddatblygu. Yn Cefndir a Nodau, rhoddwn y gwaith yn ei gyd-destun, yna trafodwn y gwaith o'i ddylunio, ei ddatblygu, a'i werthuso.

2 CEFNDIR A NODAU

Mae gofyn gael set eang o ymchwilwyr sy'n meddu ar wahanol sgiliau ar gyfer datblygu unrhyw offeryn ieithyddol. Mae'r project ColloCaid [5], [6], [7] yn dwyn ynghyd ymchwilwyr sydd â sgiliau ym maes geiriaduraeth, addysgu iaith, addysgeg, rhyngweithiad rhwng pobl a chyfrifiaduron, delweddu a rhaglennu cyfrifiaduron. Mae gan y project gyllid am bedair blynedd, oddi wrth Gyngor Ymchwil y Celfyddydau a Dyniaethau (AHRC) ac mae'n gydweithrediad rhwng academyddion yn Surrey, Bangor (DU) a Poznań, Gwlad Pwyl. Nod cyffredinol y project yw dylunio a datblygu golygydd rhyngweithiol y gallai ysgrifenwyr Saesneg Academaidd eu defnyddio i ddysgu a defnyddio cydleoiliadau. Mae'r golygydd yn caniatáu i ysgrifenwyr deipio geiriau, ac mae'r system yn amlygu ac yn rhestru cydleoiliadau. Cafodd y data sylfaenol ei ddethol yn ofalus ac yn fwriadol gan y tîm, ac mae'r rhyngwyneb dynol a'r pen blaen delweddu wedi cael ei adolygu droeon ac wedi cael ei werthuso gan ddefnyddwyr, i asesu pa mor addas ydyw i'w ddefnyddio. Gall darllenwyr roi cynnig ar yr offeryn ac archwilio gwahanol gydleoiliadau drwy fynd i wefan collocaid.uk. Mae ColloCaid yn dadansoddi'r geiriau y mae pobl yn eu hysgrifennu, yn argymhell ac yn delweddu gwahanol eiriau ar sail data systematig o gorpora testun arbenigol.

Gyda'r datblygiadau mewn ieithyddiaeth corpws, a datblygiad offer ar-lein, mae'n awr yn llawer haws i unrhyw ddefnyddiwr gael mynediad at gorpora sy'n bodoli eisoes a'u chwilio. Yn wir, mae llawer o offer ieithyddol corpora ar-lein yn cynnig i ddefnyddwyr fynediad at gorpora sydd wedi'u dadansoddi ymlaen llaw, megis y British National Corpus (BNC) a'r Contemporary of American English (COCA). Mae rhai yn caniatáu i ddefnyddwyr uwchlwytho eu dogfennau eu hunain i'w dadansoddi. Mae offer megis AntConc [8], CQPweb [9], SketchEngine [10] a Wmatrix [11] yn caniatáu i ddefnyddwyr mwy uchelgeisiol greu eu corpora arbennig eu hunain.



Ffigur 1: Gall ieithyddiaeth corpws helpu ysgrifenwyr i ddadansoddi gwahanol strwythurau mewn testun neu destunau ysgrifenedig wrth iddo gael ei greu. Gyda'r project ColloCaid ein ffocws yw helpu awduron wrth iddynt ysgrifennu a chyfansoddi eu testunau.

Ceir nifer o resymau dros ddadansoddi testun, rydym yn eu dosbarthu yn ddau grŵp ar wahân: gweler Ffigur 1. Yn gyntaf, efallai bod academyddion eisiau dadansoddi dogfennau testun a ysgrifennwyd yn flaenorol gan bobl eraill. Efallai eu bod yn gwneud hyn er mwyn deall sut i ysgrifennu dogfennau, neu destunau tebyg yn arddull awdur arall, neu efallai eu bod eisiau deall strwythur llawer o ddogfennau a deall sut mae gwahanol fersiynau o'r wybodaeth wedi newid. Yn ail, efallai bod pobl eisiau dadansoddi'r ddogfen maent yn ei hysgrifennu. Mae geiriaduron, thesawri, gwirwyr gramadeg ac yn y blaen yn helpu ysgrifenwyr i ddewis geiriau priodol a gwneud eu dogfennau'n hawdd eu deall.

Gwnaed llawer iawn o ymchwil ym maes dadansoddi dogfennau. Mae academyddion wedi defnyddio technegau ieithyddiaeth corpws i ddeall strwythur brawddegau a pharagraffau er mwyn dadansoddi teimlad neu ramadeg, dadansoddi sut mae ysgrifenwyr wedi trefnu'r ddogfen gyflawn, neu faint o gasgliadau neu fersiynau sy'n cyd-fynd â'i gilydd, neu hyd yn oed fapio sut mae gwyddoniaeth wedi addasu (Ffigur 1). Mae offer yn galluogi ymchwilwyr i ymchwilio i wahanol agweddau ar ddogfennau, o edrych ar fanylder strwythurau brawddegau i'r strwythur

cyffredinol. Gall y dadansoddi hwn helpu ymchwilwyr i ganfod sut wnaeth awduron ysgrifennu eu dogfennau (drwy ddadansoddi dogfennau), gan helpu ysgrifenyddwyr i ysgrifennu'n well, gan eu bod yn dysgu oddi wrth awduron eraill.

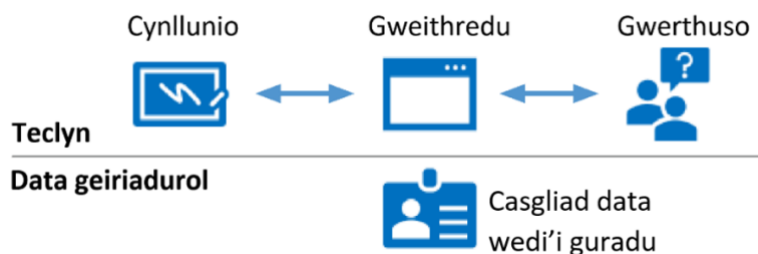
Nod ein gwaith yw helpu ysgrifenyddwyr i greu dogfennau. Defnyddiwn y term ysgrifennu agos (*close writing*) fel term cyfatebol i ddarllen agos (*close reading*). Mae darllen agos yn dechneg y mae llawer o fyfyrwyr yn gyfarwydd ag ef. Gofynnir iddynt ddadansoddi darn, cymryd nodiadau, deall llais naratif, tŷn a myfyrio'n feirniadol ar y mathau o eiriau a brawddegau y mae awdur penodol wedi'u defnyddio. Yn yr un modd, gydag ysgrifennu agos, rhaid i awduron feddwl yn ofalus am y geiriau a ddefnyddiant, rhoi eu hunain yn sefyllfa'r darlennydd a myfyrio sut caiff eu gwaith ei ystyried a'i ddeall. Yn benodol, bydd dewis cydoleoliadau priodol yn eu helpu i ysgrifennu testunau gwell a mwy dealladwy. Mae offer i helpu pobl i ysgrifennu'n well yn defnyddio technegau ieithyddol a ddefnyddir i ddadansoddi dogfennau. Gall lecsicoleg, gramadeg, dadansoddi disgwrs, arddulleg ac ati, helpu i lywio awduron i ddewis y geiriau a'r ymadroddion gorau. Caiff data (fel geiriau) eu dadansoddi a chaiff y canlyniadau eu harddangos fel gwybodaeth i hysbysu ysgrifenyddwyr. Mae'n ddull seiliedig ar batrwm, lle mae ysgrifenyddwyr yn gweld esiamplau da ac yn ddefnyddio'r wybodaeth yn eu cyd-destun eu hunain. Dysgu ar sail data (*data-driven learning*) yw hyn [12], [13], [14], [15], [16].

Heddiw, ceir llawer o offer sy'n gallu helpu pobl i ysgrifennu testunau. Mae prosesyddion geiriau, megis Word™, Apple Pages neu LibreOffice yn caniatáu i ddefnyddwyr ysgrifennu a golygu testunau. Mae golygwyr ar-lein, fel Word online gan Microsoft neu Google docs yn helpu ysgrifenyddwyr i rannu dogfennau a chael mynediad i'w testunau o bell. Mae golygwyr arbenigol yn helpu â sialensiau penodol, er enghraifft, mae OneNote neu Evernote yn caniatáu i ddefnyddwyr greu, golygu a rhannu nodiadau. Mae'r rhan fwyaf o brosesyddion geiriau heddiw yn integreiddio nodweddion golygu a fformatio – mae'r rhan fwyaf yn defnyddio dull fformatio WYSIWYG [17] (yr hyn a welwch yw'r hyn a gewch) i alluogi'r awdur i weld y fersiwn derfynol. Mae rhai prosesyddion yn gwahanu'r golygu oddi wrth y fformatio; mae golygwyr testun plaen yn caniatáu i ddefnyddwyr ysgrifennu testun ASCII, y gellir ei fformatio gan ddefnyddio ieithoedd marcio megis LaTeX, HTML neu Markdown. Mae'r rhan fwyaf o olygwyr testun yn dangos gwallau sillafu, yn gwneud awgrymiadau a gwelliannau ynglŷn â'r ramadeg. Mae golygwyr testun wedi gwella dros y blynyddoedd, o fod â nodweddion cyfyngedig i fod wedi integreiddio llawer o gymhorthion ysgrifennu. Er enghraifft, mae nodweddion lleferydd-i-destun a thestun-i-leferydd ar gael yn hawdd yn y rhan fwyaf o systemau. Mae nodweddion gwirio gramadeg wedi gwella, er enghraifft drwy integreiddio systemau megis Grammarly.

Fodd bynnag, mae llawer o ymchwil i'w wneud o hyd a cheir llawer o dechnegau y gellid eu hintegreiddio'n well â golygwyr testun. Er enghraifft, ceir nifer o systemau sy'n cyfrifo pa mor rhwydd i'w ddarllen yw testun, i broffilio testun ysgrifenedig yn erbyn rhestrau geirfa Academic Word List (AWL) [18] neu New Academic Word List (NAWL) [19], ond mae'r systemau hyn ar wahân i'r golygydd. Yn y rhan fwyaf o systemau, mae angen i'r ysgrifennwr gopïo a gludo eu testun i system ddadansoddi ar wahân i gael canlyniadau. Mae llawer o systemau yn delweddu dogfennau testun i arddangos perthnasoedd rhwng geiriau neu ddogfennau, ond unwaith eto nid ydynt wedi'u hintegreiddio'n dda â golygwyr testun. Yn olaf, ceir llawer o offer sy'n creu corpora testun pwrpasol o gyrff dogfennau, er bod y rhain eto fel arfer ar wahân i'r golygydd. Mae cael y wybodaeth mewn offer ar wahân yn golygu nad yw o anghenraid yn gyflym nac yn hawdd ei ddefnyddio. Gall fod yn ddyrws i ddysgwyr lywio ac archwilio'r wybodaeth a gyflwynir iddynt a hefyd yn anodd iddynt ddewis y geiriau gorau. Mewn geiriau eraill, mae'n aml yn anodd i ysgrifenyddwyr ysgrifennu'n feirniadol gan ddefnyddio'r offer a gynigir iddynt. Ein nod yw integreiddio ieithyddiaeth corpws, dadansoddi cydoleoliadau a delweddu geiriau mewn un golygydd testun, i alluogi ysgrifenyddwyr i ddewis y geiriau gorau, yn ogystal â dysgu fel eu bod yn dod yn well ysgrifenyddwyr.

3 METHODOLEG DATBLYGU

Wrth ddylunio a datblygu unrhyw offeryn, mae llawer o gwestiynau i'w gofyn a phenderfyniadau i'w gwneud [20]. Mae'r rhain yn cynnwys: pa mor hir yw'r cyfnod datblygu a pha adnoddau sydd ar gael? Beth yw pwrpas a gwaith yr offeryn? Ar ba ddata y mae'r offeryn yn gweithredu? Mae'n bwysig deall pwrpas yr offeryn. Heb weledigaeth glir o lle mae'r offeryn yn ffitio gydag offer eraill, byddai peidio â defnyddio'r adnoddau sydd ar gael, neu beidio â gwybod sut gellid eu defnyddio, yn gwneud y project yn anodd ei ddatblygu. Gwyddom ar ddechrau'r project yr hyn roeddem eisiau ei greu: offeryn i ganiatáu i ysgrifenyddwyr olygu'r dogfennau ac argymhell cydleoeliadau. Gwyddom hefyd am y data, a byddem yn creu, gan ddefnyddio corpora arbenigol, gronfa ddata o gydleoeliadau gydag oddeutu 30,000 o eitemau. Er bod gennym weledigaeth gyfannol i'n hofferyn gorffenedig ar y dechrau, nid oedd manylion na nodweddion penodol yr offeryn yn hysbys inni. Mae angen i ddatblygwyr allu creu offeryn a fydd yn gweithio ac a fydd yn addas at y pwrpas. Yn y newyddion ac ar y cyfryngau cymdeithasol ceir nifer o esiamplau o brojectau cyfrifiadura nad ydynt yn cael eu cwblhau a rhai nad oes modd eu defnyddio. Wedyn, mae angen gwneud gwaith cynllunio gofalus i sicrhau y bydd y project yn cael ei gyflawni yn brydlon. Yn enwedig gyda llawer o brojectau ymchwil mae'n anodd disgrifio'n union nodweddion llawn yr offeryn ar y dechrau. Gan fod syniadau'n newid ac atebion newydd yn ymddangos wrth i'r project fynd rhagddo, mae'n anodd creu manyleb gyflawn ar ddechrau project. O'r herwydd, credasom ei bod yn amhosibl creu manyleb llawn ar ddechrau'r project, a thrwy drafod bu i ni benderfynu defnyddio dull dylunio a datblygu ystwyth (agile) [21].



Ffigur 2: Yn y project dechreusom weithredu'n fuan, i greu prototeip gweithio cynnar (swyddogaethol isel). Fe wnaethom fireinio'r gweithredu drwy ailadrodd nifer o brototeipiau. Ym mhob cam, fe wnaethom brofi'r offeryn yn fewnol, a gwerthuso'r system ymysg myfyrwyr-ddefnyddwyr. Wedyn, fe wnaethom wella'r dyluniad a'r gweithrediad ym mhob cylch gwerthuso.

Creasom nifer o brototeipiau gydol y project a'u gwerthuso wrth i'r project fynd rhagddo, fel y gwelir yn Ffigur 2. Rhanasom yr offeryn yn ddwy ran: y data geiriol sylfaenol, a'r offeryn rhyngwyneb dynol. Fe wnaethom hefyd ddilyn egwyddorion ystwyth eraill. Un egwyddor ystwyth bwysig y gwnaethom ei dilyn oedd bob amser cadw fersiwn waith o'r feddalwedd. Roedd hyn yn bwysig, oherwydd ein bod eisiau cael fersiwn y gallem ei harddangos ar fyr rybudd. Yn wir, gydol y cyfnod datblygu roedd gennym nifer o fersiynau cyfochrog. Yn gyntaf, fersiwn waith gyhoeddus, gyda nodweddion yr oeddem yn barod i'w dangos yn gyhoeddus. Yn ail, fersiwn nad oedd ar gael yn gyhoeddus gyda gwell nodweddion, i'w phrofi ar ddefnyddwyr ac a fydd yn ein helpu i bwysu a mesur gwahanol atebion. Yn drydydd, fersiwn wrthi'n cael ei datblygu gyda chod na fydd ond yn rhannol gyflawn. Caiff yr egwyddorion hyn eu hegluro yn y Maniffesto Ystwyth [22], gan gynnwys: datblygu offer i fodloni cwsmeriaid drwy drosglwyddo'r feddalwedd yn gynnar ac yn barhaus; croesawu gofynion oedd yn newid hyd yn oed yn hwyr yn y

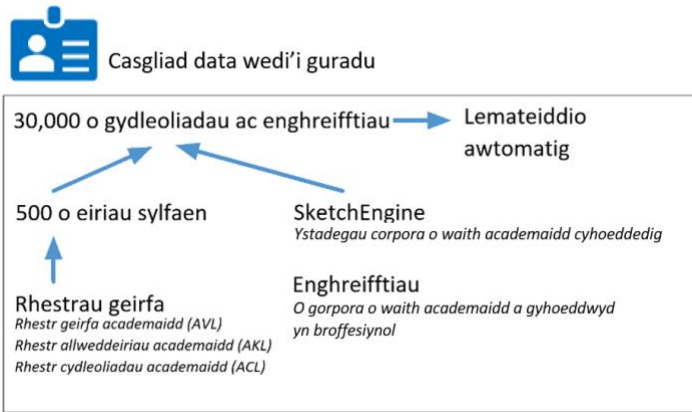
project; cyflwyno meddalwedd oedd yn gweithio yn aml; creu meddalwedd o amgylch unigolion oedd â chymhelliant cryf; cadw'r allbwn yn syml; ac adfyfrio – fel tîm – ar yr allbwn a phenderfynu sut i'w wella.

Roedd adfyfrio yn un egwyddor Ystwyth flaenllaw arall y gwnaethom ei dilyn. Fel tîm fe wnaethom ni adfyfrio'n rheolaidd ar yr offeryn. Bu i ni gynnal cyfarfodydd rheolaidd fel tîm, gan drafod yr allbwn a fersiwn gyfredol yr offeryn. Fe wnaethom brofi'r offeryn ein hunain a nodi unrhyw wallau. Tîm datblygu bychan yn unig oedd gennym, gydag ymdrechion dau ymchwilydd ôl-ddoethurol, a gan ein bod yn cadw fersiynau o'r feddalwedd a oedd yn gweithio, cafodd gwallau mawr a gwallau cystrawennol a oedd yn atal yr offeryn rhag gweithio eu datrys y tu allan i'r cyfarfodydd hyn, gan adael y cyfarfodydd i drafod gwallau semantig. Canolbwyntiodd un ymchwilydd ôl-ddoethurol ar elfennau ieithyddol a chronfa ddata'r project, gyda'r llall yn canolbwyntio ar ddatblygu'r rhyngwyneb (gweler Ffigur 2). Bu'r aelodau eraill yn helpu â'r gwaith dylunio, gwerthuso, dosbarthu, cyflwyniadau ac ati. Byddai'r tîm yn cyfarfod yn rheolaidd i drafod y dull gweithredu cyfredol, ei nodweddion, y gwallau, a gwelliannau. At hynny, fe wnaethom werthuso'r offeryn â defnyddwyr go iawn. Roeddem yn ffodus i gael mynediad at nifer o fyfyrwyr ar draws gwahanol wledydd ac o amrywiol alluoedd, o fyfyrwyr is-raddedig dechreuol i fyfyrwyr ar raglenni ieithyddol meistr uwch. Fe wnaethom werthuso'r offeryn ColloCaid ar gannoedd o ddefnyddwyr dros oes y project.

4 DYLUNIO COLLOCAID – RHYNGWYNEB DEFNYDDWYR A DATA SYLFAENOL

Fe wnaethom ddilyn un egwyddor ddylunio bwysig arall: sef symrlwydd. Er y gallem ddarparu offeryn a fyddai'n cyflwyno cannoedd o gydleoliadau posibl, fe allai ddrysu'r defnyddiwr: rhaid cael cydbwysedd rhwng nodweddion, rhwyddineb defnyddio, ac ehangder y geiriau a gynhwysir. Ni fyddai'n beth da i ysgrifennwr weld yr holl eiriau posibl ond methu â defnyddio'r offeryn. Os caiff gormod o eiriau eu harddangos, gall beri dryswch i ddefnyddwyr ynglŷn â pha un sy'n bwysig ac ansicrwydd ynglŷn â pha eiriau i'w dewis. Yn hytrach, mae angen i'r offeryn gael ei ddylunio'n ofalus er mwyn cyflwyno set addas o eiriau posibl – i roi dewis o eiriau i'r ysgrifennwr – ond fod yn hawdd ac yn glir i'w ddefnyddio.

Ein hateb oedd datblygu cronfa ddata sylfaenol o gydleoliadau a gynullwyd yn ofalus o eiriau aml eu defnydd a ganfuwyd gan ysgolheigion mewn deunydd ysgrifenedig academiaidd Prydeinig. Drwy ganolbwyntio ar y geiriau amlaf, fe lwyddon ni i ddatblygu offeryn addas i helpu ysgrifennwyr Saesneg academiaidd [6]. Mae ein cronfa ddata sylfaenol yn cynnwys dros 30 mil o eiriau wedi'u cydleoli [7]. Sefydlwyd y gronfa ddata drwy ystyried yn ofalus restrau o eiriau cyffredin gan ymchwilwyr eraill, gan gynnwys geiriau a geir yn gyffredin mewn deunydd ysgrifenedig myfyrwyr ar draws gwahanol ddisgyblaethau [23], a rhestr o eiriau allweddol academiaidd [5], [24]. Mae Ffigur 3 yn rhoi crynodeb o'r gronfa ddata sylfaenol, sy'n seiliedig ar restrau geirfaoedd, niferoedd, ac esiamplau o gorpora o ddeunydd ysgrifenedig academiaidd cyhoeddus proffesiynol, ac mae'n defnyddio lemateiddio awtomatig. Dengys Tabl 1 esiampl o'n data cydleoliadau; gyda'r gair craidd rydym yn storio'r cydleolyn (*collocate*), ynghyd â'r sgôr cysylltiad ac amllder. Mae enghraifft ddangosol i'w gweld yn y tabl.



Ffigur 3: Caiff data ColloCaid ei greu o SketchEngine [10], mae'n cynnwys nifer o gydleoliadau enghreifftiol, a chaiff ei gynnal â dros 500 o eiriau craidd sy'n deillio o restrau geirfa academaidd, cydleoliadau a geiriau allweddol.

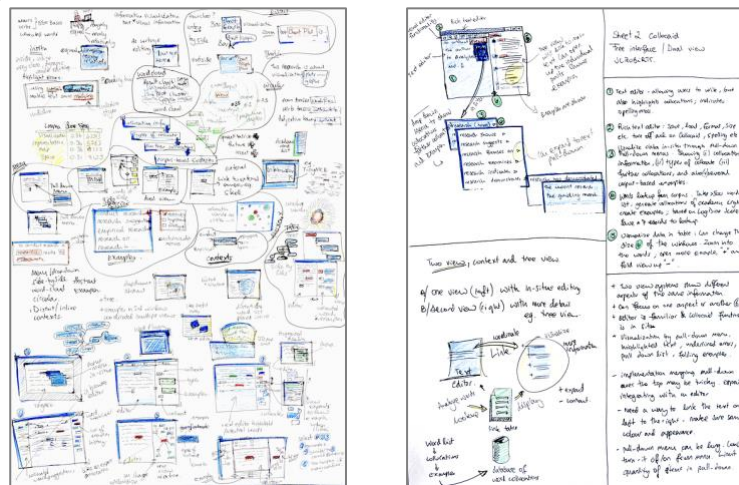
Tabl 1: Data enghreifftiol ColloCaid, yn dangos y gair craidd gyda'r cydleoliad, gwerthoedd data gwahanol ac enghraifft.

Craidd	Rhan ymadrodd	Cydleoliad Rhan ymadrodd	Cydleoliad	Amllder	Sgôr	Esiampl
access	e	b	gain	534	10.46	unauthorized users can <i>gain</i> access to it
access	e	b	ensure	113	8.09	this could also <i>ensure</i> access to information
access	e	b	restrict	114	8.66	the state has <i>restricted</i> access by foreigners
access	e	b	grant	73	8.05	the plan appeared to <i>grant</i> sole access to media for joint activities
access	e	b	secure	79	8.19	the first is <i>securing</i> universal access to justice
access	e	b	facilitate	114	8.33	technologies that might <i>facilitate</i> open access to these resources should be a priority
access	e	b	have	1818	8.49	populations who <i>have</i> access to online research
access	n	v	improve	139	8.09	<i>improving</i> access to prevent delays in answering
example	e	b	illustrates	147	10.56	this example also <i>illustrates</i> a special case of the two constraints
example	e	b	concerns	35	8.11	the next example <i>concerns</i> different interpretations
example	e	b	relates	13	6.78	the most relevant example <i>relates</i> to the manipulation of party primaries
example	e	b	inspires	13	7.59	many crimes have supposedly been <i>inspired</i> by examples shown on film
example	e	b	demonstrates	26	7.61	its nature can easily be <i>demonstrated</i> by an example
example	e	b	includes	504	10.85	examples <i>include</i> provocation and diminished responsibility
example	e	b	abounds	12	7.92	examples <i>abound</i> of some uses of the concept of gender
example	e	b	shows	110	8	an interesting example <i>shows</i> a group partition

Roedd symrlwydd hefyd yn strategaeth bwysig yr oeddem eisiau ei dilyn wrth ddiffinio'r rhyngwyneb defnyddwyr. Yn wir, mae llawer o ymchwilwyr dylunio wedi creu canllawiau dylunio ar gyfer rhyngwyneb defnyddwyr sy'n cynnwys argymhelliad i gadw pethau'n syml. Er enghraifft, rhif wyth o 10 pwynt hewristig Jakob Nilesen ar gyfer dylunio yw anelu at ddyluniad "aesthetic and minimalist" [25]. Mae'r syniad o gadw pethau'n syml, ochr yn ochr â'n dymuniad i ddylunio atebion sy'n ymarferol, yn cyd-fynd yn dda gyda'r strategaeth Ystwyth. Fodd bynnag, ceir nifer o wahanol strategaethau dylunio y gellid eu dilyn [26], a gall fod yn anodd gwybod pa un i'w dewis.

Ar y naill law, gyda'r dull Ystwyth, roeddem eisiau dechrau datblygu'r offeryn, tra ar y llaw arall, roeddem eisiau cynhyrchu astudiaeth ddylunio feirniadol ar wahân. Fel rhan o'n cynllun datblygu Ystwyth, bu i ni ddefnyddio yr holl strategaethau hyn.

Yn gyntaf, fe wnaethom greu prototeip cynnar o ColloCaid gan ddefnyddio PowerPoint™. Rhag-brototeip [27] neu frasfodel a ddefnyddiwyd gennym i ddangos y prif syniadau yw hwn. Fe wnaeth y fersiwn hon ein helpu i hyrwyddo ein hymchwili yn y cam cynnar. Fe wnaethom ei ddefnyddio mewn cyflwyniadau a'i ddangos i ymchwilwyr oedd â diddordeb. Rhoddodd hyn syniad da inni ynglŷn ag ymateb pobl. Yn wir, fe wnaeth llawer o bobl roi adborth inni gan ddweud y byddent yn ei ddefnyddio. Fe wnaeth eu hymatebion ein hysbrydoli a'n helpu i fireinio ein syniadau.



Ffigur 4: Sganiau o ddyluniadau yn dangos y ddalen gyntaf a'r ail ddalen o'r broses Pum Dalen Ddylunio. Mae'r chwith yn dangos tudalen gyda llawer o wahanol syniadau, tra bo'r dudalen dde yn dangos un dyluniad o'r offeryn terfynol sy'n dangos system dwy wedd, gyda'r golygydd testun ar y chwith a'r delweddiadau ar y dde.

Yn ail, fe wnaethom fraslunio gwahanol ddyluniadau. Roeddem eisiau ymchwilio i wahanol syniadau amgen a phwyso a mesur atebion rhyngwyneb posibl. Mae braslunio yn ddefnyddiol, oherwydd ei fod yn rhad, yn gymharol gyflym i'w greu a gellir ei wneud ar unrhyw amser. Mae'n gyflymach braslunio gwahanol atebion nag yw i'w ddatblygu a'i weithredu mewn cod. Er bod braslunio ad hoc [28] yn boblogaidd, bu i ni ddewis y fethodoleg braslunio Pum Dalen Ddylunio (FdS) i ddylunio'r rhyngwyneb [20], [29], [30]. Dewiswyd y fethodoleg FdS gan ein bod wedi'i defnyddio'n llwyddiannus mewn projectau eraill, ac mae'n darparu llif gwaith strwythuredig o ystyried nifer o syniadau dylunio gwahanol, i dair fersiwn fwy gorffenedig, a fersiwn wireddu derfynol. Bu i ni ddefnyddio dalennau papur A3, gan ddewis paled lliw llai o las, coch a melyn i uwchleuo. Dengys Ffigur 4 ddwy ddalen o'r broses hon. Dengys dalen 1 lawer o syniadau am ddyluniadau gwahanol, tra bo'r ail ddalen yn egluro mwy o fanylder un ateb penodol, gan ganolbwyntio ar y syniad o system golwg deuol cysylltiol: gyda'r golygydd testun ar y chwith a'r delweddiadau ar y dde. Pan fo defnyddiwr yn clicio ar rywbeth yn y golygydd testun, gellir ei ddangos mewn mwy o fanylder ar y llaw dde.

5 DATBLYGU A GWEITHREDU'R OFFERYN

Wrth ddatblygu unrhyw offeryn sy'n dibynnu ar olygydd testun, un o'r prif ddewisiadau i'w gwneud yw a ddylid ysgrifennu cod o egwyddorion sylfaenol, datblygu ychwanegyn ar gyfer golygydd presennol, ynteu addasu cod o

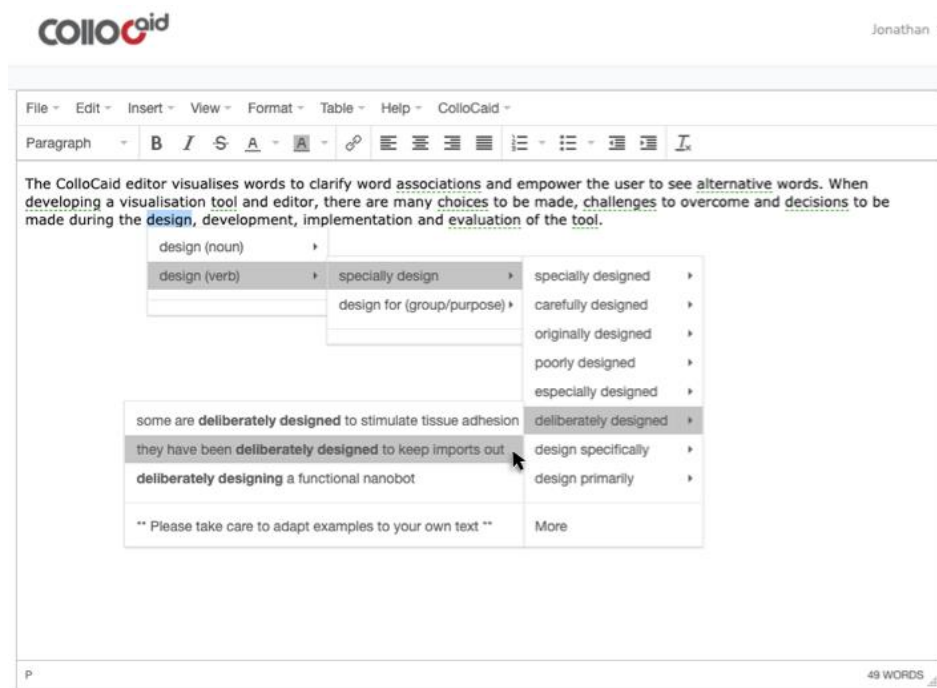
feddalwedd cod agored. Ceir manteision ac anfanteision i bob un o'r atebion hyn, ac mae angen trafod y penderfyniad yn ofalus yn y tîm datblygu gan ei fod yn newid sut caiff yr offeryn ei ddefnyddio. Bu i ni dreulio oriau lawer yn pwyso a mesur pob dull.

Byddai creu'r golygydd o ddim yn rhoi rheolaeth lawn inni dros y cod. Byddem yn gallu ei addasu i'n hanghenion, byddai'n hawdd ei ymestyn yn ddiweddarach a gallem ei ddylunio gyda nodweddion sy'n benodol i'r project. Byddai gennym hefyd hawliau llawn dros y cod a gallem ei integreiddio i unrhyw broject o'n dewis. Fodd bynnag, mae hwn yn ateb heriol sy'n gofyn am lawer o waith saernïo. Mae gofyn i'r rhaglennwr ddeall parsio, trin geiriau a nodau, golygu testun, datblygu cod rhyngwyneb defnyddwyr ac ati. Gall fod yn anodd i ddatblygwr greu cod sy'n gweithio'n briodol, heb wallau, a byddai'n sicr yn cymryd llawer o amser ac ymdrech i godio. Yn ogystal, un o'n dymuniadau oedd caniatáu i ysgrifennwyr allu copïo a gludo cod o brojectau eraill a chael nodweddion sy'n gyffredin mewn golygydd masnachol, megis newidiadau mewn ffont, penawdau, golygu lliwiau. Oherwydd yr heriau gyda chodio, dewiswyd beidio â mynd ar drywydd yr ateb hwn.

Un ateb posibl arall fyddai creu ychwanegyn. Gellir gwneud hyn mewn nifer o ffyrdd posibl. Gallai fod yn estyniad ar borwr, neu'n estyniad at brosesydd geiriau megis Microsoft Word. Yn wir, mae Microsoft Office wedi darparu gwahanol ffyrdd o ymestyn ei nodweddion. Ond ar gychwyn y project ColloCaid, roeddem yn ei chael yn anodd iawn dod o hyd i wybodaeth ac enghreifftiau o sut i ymestyn Word. Nid oedd y dogfennau ar sut i ymestyn Word yn hawdd eu deall, roedd gofyn cael llawer o wybodaeth arbenigol, ac roedd yn anodd ysgrifennu rhaglenni effeithiol. O'r herwydd, penderfynwyd peidio ag ymestyn Word. Yn benodol, nid oeddem yn gwybod yn iawn sut i integreiddio cronfa ddata, amlygu geiriau unigol, integreiddio â'r cywirydd sillafu, nac amlygu a newid geiriau. Roedd hyn cyn y gwelliannau diweddar yn nhechnoleg ychwanegion Office. Mae'r nodweddion ychwanegion newydd nawr yn caniatáu i ddatblygwyr addasu Microsoft Word, gall rhaglenwyr ddefnyddio Word JavaScript API a'r Office JavaScript API, i ymgorffori gwahanol gronfeydd data, cyflawni diogeled Single Sign-on (SSO), a defnyddio Ember, Backbone, Angular a React. Y syniad yw y gallai cod y gellid ei redeg mewn porwr gael ei redeg fel Ychwanegyn i Word, a gall rhaglenwyr greu ceisiadau i weithredu ar wrthrychau Word a chydamseru cyflwr gwrthrychau.

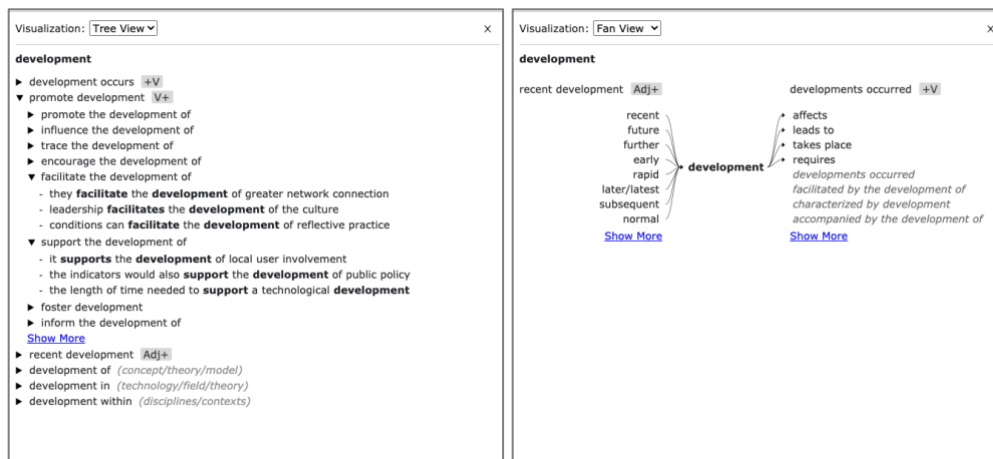
Y trydydd ateb yw addasu golygydd cod agored. Mae yna lawer o olygwyr y gallem eu dewis. Mae gan gymwysidiadau bwrdd gwaith fel Atom gymuned o ddatblygwyr, trefniant pecynnau GitHub a chymorth ar draws llwyfannau. Yn yr un modd, mae'r golygydd Brackets yn cynnig cymorth ar draws llwyfannau a chymuned GitHub. Dewis arall yw datblygu ar-lein. Golygwyr HTML sy'n darparu amgylchedd WYSIWYG ar-lein yw meddalwedd megis Quill, EditorJS a TinyMCE. Ar ôl ystyried, penderfynwyd ddatblygu offeryn ar-lein, a dewiswyd y golygydd TinyMCE, sy'n defnyddio Trwydded Gyhoeddus Gyffredinol GNU. Fodd bynnag, roedd heriau yn wynebu ein penderfyniad i ddefnyddio TinyMCE o hyd: er mwyn ei ymestyn yn iawn, roedd arnom angen deall sut mae'r TinyMCE yn gweithio, a sut yr ysgrifennwyd y cod. Er ein bod wedi treulio llawer o amser yn goresgyn yr heriau hyn, llwyddasom i wneud llawer o ychwanegiadau i'r cod, ac ar ôl ychwanegu nodwedd newydd, fe wnaethom ei phrofi'n drylwyr.

Fe wnaethom ymestyn TinyMCE i ddarparu'r nodweddion yr oedd arnom eu hangen. Mae Ffigur 5 yn dangos sgrinlun o'r golygydd. Rydym yn parsio'r geiriau sydd wedi'u hysgrifennu, gan eu harchwilio yn ein cronfa ddata cydoleoliadau. Dangosir y geiriau yn y gronfa ddata i'r ysgrifennwr wedi'u tanlinellu â llinell doredig werdd. Ychwanegwyd cwymplenni in-situ sy'n ymddangos pan fo ysgrifennwr yn clicio ar air wedi'i danlinellu. Mae'r cwymplenni yn caniatáu i ysgrifennwyr ddewis gwahanol rhannau ymadrodd, cydoleoliadau penodol ac enghreifftiau. Gall ysgrifennwyr glicio ar unrhyw rai o'r geiriau yn y ddewislen i'w mewnosod yn y ddogfen.



Ffigur 5: Y golygydd ColloCaid yn dangos y cwymplenni. Gall ysgrifenydd wneud dewisiadau ynglŷn â'r data a welant, a chaiff unrhyw elfen a ddewisant ei mewnosod yn y testun.

Rydym wedi dylunio nifer o ddelweddiadau amgen. Fe wnaethom ddylunio tablau, rhestrau, delweddu cylch ac arddulliau eraill. Rydym wedi gweithredu rhai o'r syniadau hyn. Mae Ffigur 6 yn dangos y 'Tree View' a'r 'Fan View'. Gyda'r 'Tree View' (panel chwith) gall defnyddwyr ddewis yr eiconau triongl i ddatgelu manylion am y cydoleoliadau a dangos brawddegau enghreifftiol. Fel gyda'n holl delweddu, pan mae'r defnyddiwr yn clicio ar linell, caiff y testun ei osod yn y ddogfen. Mae'r panel dde yn Ffigur 6 yn dangos sgrinlun o'r 'Fan View', sy'n dangos pa gydoleoliadau y gellir eu defnyddio cyn, neu ar ôl, y gair a ddewiswyd. Yn yr holl ddelweddu, pan fo'r ysgrifennwr yn dewis "Show More", cynhwysir rhestr o eiriau ychwanegol. Pan fo'r rhestr yn hirach na'r sgrin, ychwanegir bar sgrollo. Hefyd, gellir dangos rhagor o wybodaeth ar gyfer pob un o'r geiriau hyn, a darparwn ddolen uniongyrchol at restr sy'n dangos y patrwm geiriau a ddewiswyd yn SkELL (Sketch Engine for Language Learning). Offeryn i helpu pobl i archwilio sut caiff termau penodol eu defnyddio gan siaradwyr Saesneg go iawn yw SkELL, ac mae'n rhoi esiamplau da o gydoleoliadau a chyfystyron a gafwyd yn awtomatig o gorpws mawr.



Ffigur 6: Dau ddelweddiad yn yr offeryn ColloCaid. Mae'r panel ar y chwith yn dangos y 'Tree View', lle gall defnyddwyr glicio ar y trionglau i ddangos agor neu gau gwybodaeth ychwanegol. Dengys y panel ar y dde sgrinlun o'r delweddiadau yn y 'Fan View' ar y gair *development*. Mae hwn yn darlunio geiriau a gydeolir y gellir eu rhoi cyn, neu ar ôl, y term a ddewiswyd. Caiff rhestrau o eiriau ychwanegol eu dangos pan fo'r ysgrifennwr yn dewis 'show more'.

6 GWERTHUSO AC ADBORTH

Mae yna sawl ffordd o werthuso offer [31]. Roedd angen inni ddilyn strategaeth werthuso syml, a hawdd ei chyflawni, a fyddai'n cyd-fynd â'r fethodoleg datblygu Ystwyth. Roedden ni hefyd eisiau dull y gallem ei ddefnyddio gyda myfyrwyr, a fyddai'n hawdd ei ddeall mewn gwahanol ieithoedd. Gallem ofyn cwestiynau fel a fyddai'r ysgrifennwr yn defnyddio'r offeryn eto? Gallem ofyn cwestiynau am ba mor effeithlon, hawdd dysgu ei ddefnyddio, atyniadol ac ati ydyw, a pha mor hawdd ei ddefnyddio, gan greu ein set ni ein hunain o gwestiynau at y pwrpas. Fodd bynnag, penderfynwyd ddilyn y Raddfa Defnyddioldeb Systemau (SUS) [32]. Mae'r SUS yn cynnwys deg cwestiwn syml, y mae cyfranogwyr yn eu hateb ar raddfa 5 pwynt Likert, o "anghytuno'n gryf" i "cytuno'n gryf".

1. Rwy'n meddwl yr hoffwn ddefnyddio'r system hon yn aml.
2. Roedd y system yn ddiangen o gymhleth yn fy marn i.
3. Roeddwn i'n meddwl bod y system yn hawdd ei defnyddio.
4. Rwy'n meddwl y byddai angen imi gael cymorth unigolyn technegol i fod yn gallu defnyddio'r system hon.
5. Rwy'n teimlo bod yr amrywiol nodweddion yn y system hon wedi'u hintegreiddio'n dda.
6. Roeddwn i'n meddwl bod gormod o anghysondeb yn y system hon.
7. Byddwn yn dychmygu y byddai'r rhan fwyaf o bobl yn dysgu i ddefnyddio'r system hon yn gyflym iawn.
8. Roeddwn i'n teimlo bod y system yn lletchwith iawn i'w defnyddio.
9. Roeddwn i'n teimlo'n hyderus iawn yn defnyddio'r system.
10. Roedd angen imi ddysgu llawer o bethau cyn y gallwn fynd ati i ddefnyddio'r system hon.

Gan ddefnyddio'r dull hwn, gallwn gasglu data meintiol ar ddefnyddioldeb y system ymysg defnyddwyr. Hefyd, penderfynwyd ychwanegu dau gwestiwn penagored arall, yn gofyn i gyfranogwyr egluro rhywbeth roeddent yn ei hoffi, a rhywbeth arall nad oeddent yn ei hoffi, am yr offeryn. Fe wnaethom ddefnyddio'r SUS gan fod y deg cwestiwn wedi'u cyfieithu i lawer o ieithoedd (ac felly roedden ni'n gallu defnyddio'r cyfieithiadau, megis gyda'n profwyr

Pwylaidd), rydym wedi'i ddefnyddio mewn projectau eraill, a gellir gweinyddu'r deg cwestiwn yn gyflym i'r cyfranogwyr heb fod angen esboniadau hirfaith.

Defnyddiwyd yr holiadur hwn i werthuso'r offeryn ar wahanol gyfnodau o'i ddatblygiad. Gwerthuswyd yr offeryn gyda dros 140 o gyfranogwyr ar draws pum safle: León, Paris, Porto Alegre, São José do Rio Preto, a Poznań, a chawsom lawer o adborth gan y defnyddwyr. Er enghraifft, roedd gan un o'n prototeipiau cynnar dabl ar waelod ffenestr y golygydd a ddangosai restr gyflawn o gydleoliadau o'r testun cyfredol. Byddai'r tabl yn diweddar wrth i'r ysgrifennwr fynd yn ei flaen. Fodd bynnag, drwy'r gwerthuso, dywedodd y defnyddwyr fod y testun yn "dawnsio" o gwmpas, ac nad oedd safle'r tabl (islaw'r golygydd) yn ddefnyddiol. Gan ymateb i'r adborth hwn, fe wnaethom ddileu'r 'Table View', ac yn hytrach defnyddio'r 'Tree View' (gweler Ffigur 6).

7 CRYNODEB A CHASGLIADAU

Defnyddiwyd methodoleg datblygu Ystwyth gennym a bu inni ddatblygu'r offeryn yn barhaus o brototeipiau cynnar i'r fersiwn orffenedig ac rydym wedi gwerthuso ColloCaid gyda defnyddwyr go iawn. Rydym yn crynhoi pum elfen a wnaeth ein helpu i lwyddo. I ddechrau, treuliwyd llawer o amser ac ymdrech gennym yn cydgasglu'r data cydleoliadau sylfaenol. Mae'n bwysig creu data cywir, heb ddata cydleoliadau ac enghreifftiau go iawn, ni fyddai'r offeryn yn ddefnyddiol. Yn ail, un o'r agweddau pwysig ar ein dull Ystwyth oedd cadw dwy fersiwn a oedd yn gweithio o'r offeryn ac un fersiwn a oedd wrthi'n cael ei datblygu. Golygai hyn ein bod yn gallu arddangos a chyflwyno ein gwaith ar unrhyw adeg yn ystod y project. Yn drydydd, drwy ymestyn y golygydd TinyMCE roeddem yn gallu datblygu offeryn sy'n gweithredu mewn ffordd draddodiadol, gydag ysgrifennwyr yn gweithio ar ryngwyneb cyfarwydd. Yn bedwerydd, drwy ddylunio, datblygu a gwerthuso gwahanol syniadau delweddu, a'n pwyslais ar symrlwydd, llwyddasom i greu delweddu defnyddiol. Yn olaf, fe wnaeth ein rhaglen o werthuso a phrofi parhaus, wedi'i hintegreiddio â'r Raddfa Defnyddioldeb Systemau (SUS) ein galluogi i gael adborth defnyddiol a phwysig ar yr offeryn sydd wedi ein helpu i'w wella'n barhaus.

I gloi, rydym wedi cyflwyno'r golygydd ColloCaid, sy'n offeryn ysgrifennu i helpu ysgrifennwyr wrth iddynt olygu testunau. Mae'r offeryn yn awgrymu cydleoliadau naturiol ar gyfer geiriau pan mae'r defnyddiwr yn eu teipio ac yn delweddu geiriau i ddarparu awgrymiadau amgen.

DIOLCIADAU

Ariennir y gwaith hwn drwy grant AH/P003508/1Cyngor Ymchwil y Celfyddydau a Dyniaethau (AHRC), <http://collocaid.uk>

CYFEIRIADAU

- [1] Ramesh Krishnamurthy. 1987. The process of compilation. Yn Looking up: An account of the COBUILD project in lexical computing. Collins ELT, Llundain a Glasgow, 62-85.
- [2] Jonathan C. Roberts, H. Al-Manee, Peter Butcher, Robert Lew, Geraint Rees, Nirwan Sharma ac Ana Frankenberg-Garcia. 2019. Multiple Views: different meanings and collocated words. Computer Graphics Forum 38, 3 (Mehefin 2019), 79-93. 10.1111/cgf.13673
- [3] Alex Boulton. 2016. Integrating corpus tools and techniques in ESP courses. Asp 69, 69 (Mawrth 2016), 113-137. 10.4000/asp.4826
- [4] Ana Frankenberg-Garcia. 2018. Investigating the collocations available to EAP writers. Journal of English for Academic Purposes 35 (Gorffennaf 2018), 93-104.
- [5] Kirsten Ackermann ac Yu-Hua Chen. 2013. Developing the Academic Collocation List (ACL)--A corpus-driven and expert-judged approach. Journal of English for Academic Purposes 12, 4 (Rhagfyr 2014), 235-247.
- [6] Ana Frankenberg-Garcia, Robert Lew, Jonathan C. Roberts, Geraint Paul Rees a Nirwan Sharma. 2019. Developing a writing assistant to help EAP writers with collocations in real time. ReCALL 31, 1 (Ionawr 2019), 23-39.
- [7] Jonathan C. Roberts, Peter W. S. Butcher, Robert Lew, Geraint Paul Rees, Nirwan Sharma ac Ana Frankenberg-Garcia. 2020. Visualising

Collocation for Close Writing. Yn EuroVis 2020 - Short Papers. The Eurographics Association, Ar-lein.

- [8] Laurence Anthony. 2018. Visualisation in corpus-based discourse studies. Yn *Corpus Approaches To Discourse*. Routledge. Llundain. 197-224.
- [9] Andrew Hardie. 2012. CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17, 3 (Ionawr 2012), 380-409.
- [10] Adam Kilgariff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý a Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1, (Gorffennaf 2014), 7-36.
- [11] Paul Rayson. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13, 4 (Ionawr 2008), 519-549.
- [12] Tim Johns. 1991. Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal* 4, (1991), 1-16.
- [13] Ana Frankenberg-Garcia. 2012. Raising teachers' awareness of corpora. *Language Teaching* 45, 4 (Hydref 2012), 475-489.
- [14] Ana Frankenberg-Garcia. 2020. Combining user needs, lexicographic data and digital writing environments. *Language Teaching* 53, 1 (Ionawr 2020), 29-43.
- [15] Ana Frankenberg-Garcia. 2016. Corpora in ELT. Yn *The Routledge Handbook of English Language Teaching*. Routledge. Llundain. 401-416. 10.4324/9781315676203
- [16] Alex Boulton a Tom Cobb. 2017. Corpus Use in Language Learning: A Meta-Analysis. *Language Learning* 67, 2 (Rhagfyr 2017), 348-393.
- [17] Brad Myers. 1998. A brief history of human-computer interaction technology. *Interactions* 5, 2 (Mawrth/Ebrill 1998), 44-54. 10.1145/274430.274436,2,44-54
- [18] Dee Gardner a Mark Davies. 2014. A new academic vocabulary list. *Applied linguistics* 35, 3 (Gorffennaf 2014), 305-327.
- [19] Averil Coxhead. 2000. A New Academic Word List. *TESOL Quarterly* 34, 2 (Haf 2000), 213-238. <https://doi.org/10.2307/3587951>
- [20] Jonathan C. Roberts, Chris Headleand a Panagiotis D. Ritsos. 2016. Sketching Designs Using the Five Design-Sheet Methodology. *Transactions on Visualisation and Computer Graphics* 22, 1 (Ionawr 2016), 419-428.
- [21] Barry Boehm. 2002. Get ready for agile methods, with care. *Computer* 35, 1 (Ionawr 2002), 64-69. 10.1109/2.976920
- [22] Martin Fowler a Jim Highsmith. 2001. The agile manifesto. *Software Development* 9, 8 (2001), 28-35.
- [23] Sian Alsop a Hilary Nesi. 2009. Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora* 4, 1 (Mai 2009), 71-83.
- [24] Magali Paquot. 2010. Academic vocabulary in learner writing: From extraction to analysis. Continuum. Llundain, Efrog Newydd.
- [25] Jakob Nielson. 1994, 2020. 10 Usability Heuristics for User Interface Design. <https://www.nngroup.com/articles/ten-usability-heuristics/>
- [26] Alex Bigelow, Steven Drucker, Danyel Fisher a Miriah Meyer. 2014. Reflections on How Designers Design with Data. Yn *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. Association for Computing Machinery, Efrog Newydd, NY, UDA, 17-24.
- [27] Marc Rettig. 1994. Prototyping for Tiny Fingers. *Communications of the ACM* 37, 4 (Ebrill 1994), 21-27.
- [28] Bill Buxton. 2010. Sketching user experiences: getting the design right and the right design. Morgan Kaufmann.
- [29] Jonathan C. Roberts, Christopher J. Headleand a Panagiotis D. Ritsos. 2017. Five Design-Sheets -- Creative design and sketching in Computing and Visualization. Springer International Publishing, Y Swistir.
- [30] Jonathan C. Roberts. 2011. The Five Design-Sheet (FdS) approach for Sketching Information Visualization Designs. Yn *Proceedings of Eurographics 2011 - Education Papers*. The Eurographics Association, Llandudno, Cymru, 27-41.
- [31] Jeff Sauro. 2015. SUPR-Q: a comprehensive measure of the quality of the website user experience. *Journal of Usability Studies* 10, 2 (Chwefror 2015), pp.68-86. 10.5555/2817315.2817317
- [32] Aaron Bangor, Philip T. Kortum a James T. Miller. 2008. An empirical evaluation of the system usability scale. *International Journal of Human Computer Interaction* 24, 6 (Gorffennaf 2008), 574-594.

Modelu a Phrosesu Testun Cymraeg gan ddefnyddio Pecyn Cymorth TAWA

LEENA SARAH FARHAT

Ysgol Gwyddorau Cyfrifiadurol a Pheirianeg Electronig, Prifysgol Bangor, Cymru

WILLIAM JOHN TEAHAN

Ysgol Gwyddorau Cyfrifiadurol a Pheirianeg Electronig, Prifysgol Bangor, Cymru

Mae'r bennod hon yn trafod canlyniadau rhagarweiniol cymhwyso'r pecyn cymorth NLP ar sail cywasgu TAWA i amrywiol dasgau NLP Cymraeg gan ddefnyddio corpora CEG, CorCenCC ac UAGT-PNAW.

Geiriau ac Ymadroddion allweddol: Pecyn Cymorth TAWA, modelu iaith ar sail cywasgu, CEG, CorCenCC, UAGT-PNAW

1 CEFNDIR A CHYMHELLIANT

Mae'r bennod hon yn disgrifio ymchwiliad rhagarweiniol i gymhwyso'r pecyn cymorth TAWA [1] i'r broblem o fodelu a phrosesu testun Cymraeg.

Mae'r pecyn cymorth *Text Analysis from Waikato*, a elwir hefyd yn TAWA, yn becyn cymorth sy'n defnyddio modelau iaith ar sail cywasgu ar gyfer dadansoddi testun, cloddio testun a phrosesu iaith naturiol (NLP). Mae'r pecyn hwn wedi bod yn cael ei ddatblygu'n barhaus ers dros 20 mlynedd. Wedi'i roi ar waith yn C yn wreiddiol, mae ganddo bellach amrywiaeth o adnoddau hawdd eu defnyddio yn ogystal â llyfrgelloedd mwy pwerus sy'n cael eu datblygu yn Python.

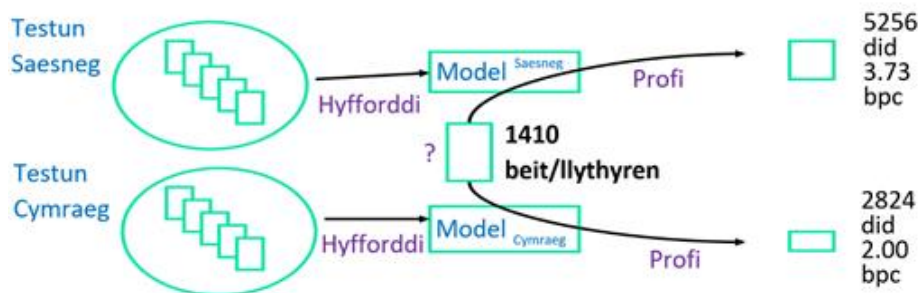
Mae'r pecyn cymorth wedi'i gymhwyso'n llwyddiannus i ystod eang o dasgau dadansoddi testun, cloddio testun ac NLP ac mae'n aml yn cynhyrchu canlyniadau o'r radd flaenaf mewn llawer o'r meysydd hyn [1]. Mae rhai rhaglenni enghreifftiol yn cynnwys: Alinio brawddegau dwyieithog Saesneg-Tsieineaidd a Saesneg-Arabeg; cryptoddadansoddi seiffrau trawsddodi; dosbarthu a segmentu testun Arabeg; adnabod emosiwn mewn testun Saesneg ac Arabeg; categorioedd rhywedd ac awdur trydariadau Arabeg; a chanfod swyddogaeth genynnau mewn cyhoeddiadau bioleg.

Mae cwestiwn yn codi ynghylch pa mor llwyddiannus fyddai'r pecyn cymorth o'i gymhwyso i'r iaith Gymraeg. Y prif gymhelliad y tu ôl i'r bennod hon yw i gynhyrchu gwaith rhagarweiniol sy'n ceisio ateb y cwestiwn hwn.

1.1 Modelu iaith ar sail Cywasgu

Ar hyn o bryd mae Pecyn Cymorth TAWA yn defnyddio'r algorithm cywasgu Rhagfynegi drwy Gyfatebu'n Rhannol (PPM) [2] sydd wedi bod yn arloesol o ran algorithmau cywasgu ers dros dri degawd ac sy'n parhau i fod yn un o'r rhaglenni cywasgu di-golled gorau ar gyfer testunau iaith naturiol. Mae'r defnydd o gywasgu ar gyfer Prosesu Iaith

Naturiol (NLP) o ddiddordeb oherwydd bod ganddo'r galli i gael gwared â data diangen ac mae modd dilysu unrhyw fodel sy'n cael ei greu. Mae modd gwirio unrhyw rhagfynegiad mae model yn ei wneud drwy ei amgodio ac yna ei ddadgodio. Gall TAWA ddosbarthu testun gan ddefnyddio Lleiafswm Trawsentropi. Mae hyn yn defnyddio hyd cod cywasgiad (nifer y didau sydd eu hangen i amgodio llinyn testun), sy'n cael ei weld fel trawsentropi yn yr achos hwn, fel ffordd o fesur "daioni" y model iaith sydd wedi'i greu.



Ffigur 1: Darlun gweledol o sut mae cywasgu'n cael ei ddefnyddio yn TAWA ar gyfer y dasg adnabod iaith. Mae maint llai yr allbwn wedi'i gywasgu ar gyfer y model Cymraeg ar y testun prawf yn dangos mai'r Gymraeg yw'r iaith fwyaf tebygol.

Mae Ffigur 1 yn cynnig darlun gweledol o sut mae hyn yn gweithio. Mae'r diagram yn dangos dau fodel a grëwyd drwy eu hyfforddi ar destun sy'n cynrychioli dwy iaith – Cymraeg a Saesneg – gan ddefnyddio'r adnodd hyfforddi a ddarperir gan y pecyn cymorth. Defnyddir y modelau hyn wedyn gan yr offeryn dosbarthu i gywasgu testun prawf penodol, y mae ei faint heb ei gywasgu yn 1410 nod (fel sy'n cael ei ddangos yn y ffigur). Mae'r offeryn yn canfod mai'r model Cymraeg sy'n cywasgu'r testun profi orau, gan ddefnyddio dim ond 2824 did, neu 2.00 did i bob nod ("bpc"). Mae hon yn gyfradd cywasgu sylweddol, ac yn ganlyniad cyffredin gyda thestunau iaith naturiol. Mae hefyd yn llawer gwell na'r model Saesneg wrth gywasgu'r testun (5256 did neu 3.73 bpc). Unwaith eto, mae'r gwahaniaeth hwn o ran cywasgu rhwng y model Saesneg a'r model Cymraeg yn nodweddiadol. Mae hyn yn awgrymu'n gryf mai Cymraeg yw iaith y testun prawf.

Mae TAWA hefyd yn defnyddio pensaernïaeth model sianel ddi-sŵn ar gyfer rhaglenni sy'n seiliedig ar gywiriadau, drwy eu hail-lunio fel math o broblem mewnosod tagiau [1]. Mae hyn i gyd yn cael ei ddwyn ynghyd yn y pecyn cymorth drwy sicrhau bod offer hawdd i'w defnyddio sy'n seiliedig ar gywiriadau ar gael i ddefnyddwyr eu defnyddio gyda thasgau NLP amrywiol fel segmentu a mewnosod tagiau.

Mae'r ystod o raglenni a ddarperir yn y pecyn cymorth i'w weld yn nhabl 1.

Tabl 1: Yr ystod o raglenni a ddarperir ym mhhecyn cymorth TAWA

Rhaglen	Disgrifiad
align	Rhaglen ar sail cywasgu ar gyfer dilysu aliniad y brawddegau mewn corpws cyfochrog.
classify	Offeryn dosbarthu testun ar gyfer dod o hyd i'r model hyfforddi sy'n cywasgu'r testun prawf orau.
codelength	Offeryn ar gyfer cyfrifo cywasgiad hyd cod heb godio y gellir ei ddefnyddio i ragweld effeithlonrwydd cywasgiad penodol.
decode	Offeryn sy'n datgywasgu'r ffeil mewnbwn gywasgedig o ffeil ffisegol ar ddisg.
encode	Offeryn sy'n cywasgu ffeil mewnbwn, gan ysgrifennu'r allbwn cywasgedig i ffeil ffisegol ar ddisg.
markup	Offeryn sy'n defnyddio dull sy'n seiliedig ar gywasgu i anodi ffeil testun drwy fewnosod tagiau'n awtomatig (e.e. ar gyfer cyfnewid cod, gan farcio lle mae'r ieithoedd penodol yn digwydd yn y testun).
segment	Rhaglen cywasgu a ddefnyddir i segmentu geiriau mewn ffeil testun.

Rhaglen	Disgrifiad
train	Rhaglen a ddefnyddir i adeiladu modelau iaith sy'n seiliedig ar gywasgu drwy eu "hyfforddi" ar destun. Defnyddir y modelau hyn wedyn gan y rhaglenni eraill yn y pecyn cymorth. Mae dau brif fath o fodel yn bosibl: modelau dynamig, sy'n cael eu diwedddaru wrth i'r rhaglen fynd yn ei blaen (a ddefnyddir yn bennaf ar gyfer gweithrediadau amgodio a dadgodio addasol); a modelau statig, nad ydynt yn cael eu diwedddaru ar ôl iddynt gael eu creu.

Mae strwythur gweddill y bennod hon fel a ganlyn. Bydd adran 2 yn trafod y tri chorpws Cymraeg a ddefnyddir yn y gwerthusiad arbrofol a drafodir yn adran 3. Mae'r arbrofion yn ymchwilio i ddefnydd y pecyn cymorth TAWA ar gyfer y tasgau NLP canlynol: cywasgu; dosbarthu testun; ac alinio brawddegau ar gyfer corpora cyfochrog. Mae'r adran olaf yn darparu trafodaeth ac yn crybwyll gwaith sydd i ddod yn y dyfodol.

2 CORPORA ARBROFOL

Er mwyn defnyddio TAWA yn effeithiol, mae'n aml yn werth defnyddio corpora testun i hyfforddi'r modelau iaith. Ar gyfer yr astudiaeth hon sy'n ymwneud â'r Gymraeg, mae hyn yn cyfyngu ar ein dewisiadau gan ei bod yn iaith lai ei hadnoddau. O ganlyniad i hyn, dau ddewis posibl oedd y corpws CEG [3] a'r corpws CorCenCC mwy diweddar [4], [5]. Mae'r ddau wedi'u cynllunio fel corpora cytbwys at sawl diben er mwyn darparu enghreifftiau o ddefnydd yr iaith. Mae'r trydydd corpws ar y llaw arall, corpws UAGT-PNAW [6], yn gorpws cyfochrog sy'n cynnwys testun Cymraeg a Saesneg. Ychwanegwyd y corpws hwn at y gwerthusiad arbrofol er mwyn penderfynu pa mor effeithiol yw offeryn alinio brawddegau TAWA.

Ceir disgrifiad pellach o'r tri chorpws isod.

2.1 Corpws CEG (1,046,551 o eiriau; 6,169,422 o nodau)

Mae'r corpws 6.2 MB, Cronfa Electroneg o Gymraeg (CEG), yn gasgliad rhyngddisgyblaethol o dros 1 miliwn o eiriau Cymraeg a gasglwyd o 500 testun enghreifftiol. Mae pob testun yn cynnwys 2000 gair o ryddiaith gyfoes amrywiol. Mae'r samplau'n cynnwys deunyddiau o nofelau a straeon byrion, ysgrifennu crefyddol, llenyddiaeth plant, darlithoedd cyhoeddus, papurau newydd a chylchgronau, atgofion, ysgrifennu academaidd, darnau gweinyddol cyffredinol, amrywiol ddeunyddiau ym meysydd addysg, gwyddoniaeth, busnes, gweithgareddau hamdden a mwy. Cafodd y corpws hwn ei ddefnyddio i greu a chynnal CySill [3].

2.2 Corpws CorCenCC (13,443,584 o eiriau; 65,478,526 o nodau)

Mae'r Corpws Cenedlaethol o Gymraeg Cyfoes (CorCenCC) 65.4 MB yn brosiect rhyngddisgyblaethol ac aml-sefydliadol diweddar sydd wedi creu corpws cod agored o Gymraeg cyfoes ar raddfa fawr. Mae'r corpws hwn yn cynnwys enghreifftiau o iaith lafar, ysgrifenedig a/neu e-iaith o gyd-destunau bywyd go iawn. Y corpws Cymraeg cyntaf o'r math yma, mae'n cynnig darlun manwl gywir o'r Gymraeg ar draws set amrywiol o gyd-destunau defnydd. Mae'r rhain yn cynnwys negeseuon sgwrsio preifat, cymdeithasu mewn grŵp, cyfarfodydd busnes a sefyllfaoedd gwaith eraill, addysg, amrywiol gyfryngau, a sgysiau mewn llefydd cyhoeddus. Mae'n cynnwys enghreifftiau o benawdau newyddion, negeseuon e-bost a gohebiaeth bersonol a phroffesiynol, ysgrifennu academaidd, areithiau ffurfiol ac anffurfiol, negeseuon blog a negeseuon testun. Casglwyd y data iaith gan amrywiaeth o siaradwyr Cymraeg, sy'n cynrychioli pob rhanbarth yng Nghymru, yn ogystal â phob oed, rhyw a galwedigaeth, ac o amrywiaeth o gefndiroedd ieithyddol, er mwyn cynrychioli'n gywir yr ystod o fathau o destun a siaradwyr Cymraeg a geir yng Nghymru heddiw [4], [5].

2.3 Corpws UAGT-PNAW (510,813 pâr o frawddegau wedi'u halinio)

Mae'r corpws hwn yn gorpws dwyieithog 100 MB Cymraeg-Saesneg o Gofnod Cynulliad Cenedlaethol Cymru [6]. Mae'n darparu set o destun wedi'i gyfieithu ar gyfer y Gymraeg a'r Saesneg wedi'u trefnu fel parau o frawddegau wedi'u halinio.

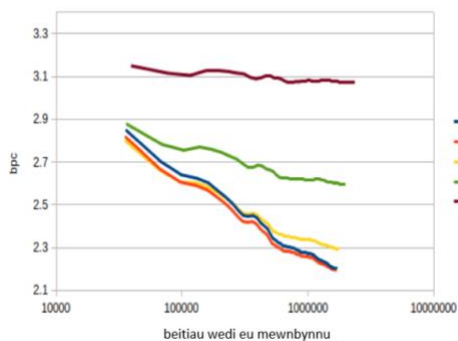
3 CANLYNIADAU ARBROFOL

Gwnaed nifer o arbrofion gan ddefnyddio rhaglenni TAWA a'r tri chorpws a ddisgrifiwyd uchod i ganfod y ffordd fwyaf effeithiol o fodelu a phrosesu testun Cymraeg wrth ddefnyddio modelau iaith sy'n seiliedig ar gywasgu. Cychwynnodd y set gyntaf o arbrofion drwy ymchwilio i ba fodelau a chorpora hyfforddi oedd y mwyaf effeithiol o ran cywasgu. Ar ôl penderfynu ar hynny, defnyddiwyd y wybodaeth hon gyda rhaglenni eraill yn yr arbrofion a ddaeth wedyn.

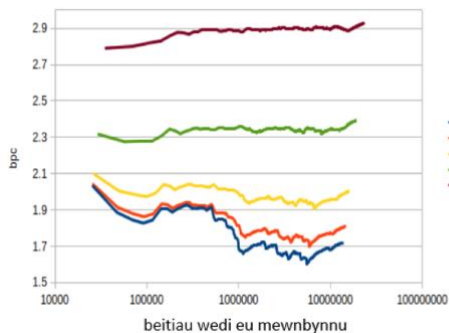
3.1 Cywasgu Testun Cymraeg gan Ddefnyddio Modelau Dynamig

Roedd y set gyntaf o arbrofion yn cynnwys cywasgu testun Cymraeg mewn modd addasol yn unig drwy ddewis modelau dynamig gyda'r offeryn **encode** yn TAWA. Ar gyfer y modelau hyn, mae'r amcangyfrifon tebygolrwydd symbolau yn wag ar ddechrau'r broses o gywasgu testun, ac wedi hynny maent yn cael eu diweddaru wrth i'r broses cywasgu fynd rhagddi. Yn aml gall modelau dynamig berfformio'n well na modelau statig gan fod PPM yn gallu addasu'n gyflym i'r testun penodol sy'n cael ei gywasgu, ond nid oes gan fodelau statig y gallu i newid yr amcangyfrifon tebygolrwydd gan fod y testun yn cael ei brosesu'n ddilysniannol.

Mae'r arbrofion cychwynnol hyn yn hanfodol i ganfod pa fodelau sydd fwyaf effeithiol ar gyfer y Gymraeg fel y gellir eu defnyddio wedyn i gynhyrchu'r canlyniadau gorau ar gyfer y cymwysiaid eraill. Yn y set gyntaf o arbrofion, cafodd y ddau gorpws testun, CEG a CorCenCC, eu cywasgu'n uniongyrchol ar gyfer modelau PPMD [1] â threfnau gwahanol, o drefn 6 i drefn 2. Mae'r canlyniadau i'w gweld yn Ffigurau 2 a 3. Mae'r graffiau'n plotio'r gymhareb cywasgu (bpc ar gyfer didau fesul nod) o'i gymharu â mewnbwn y beitiau ar raddfa log ac yn datgelu pa mor dda y mae'r cywasgu'n mynd wrth i'r ffeiliau testun gael eu prosesu.



Ffigur 2: Amgodio'r corpws CEG yn ddynamig



Ffigur 3: Amgodio corpws CorCenCC yn ddynamig.

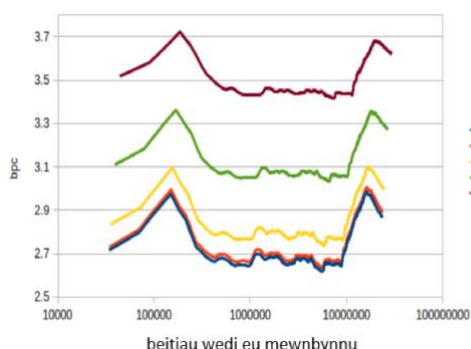
Mae'r canlyniadau'n dangos bod cywasgu'n gwella'n gyson ar gyfer y corpws CEG wrth i'r modelau gael eu diweddaru gan ychwanegu mwy a mwy o destun. Mae hwn yn ganlyniad disgwylidig ar gyfer testun iaith naturiol ac mae'n dangos bod gan yr iaith nodweddion cyson ar y cyfan h.y. nid yw'r math o iaith yn amrywio'n sylweddol. Mae gwelliannau sylweddol o ran cywasgu o fodelau trefn 2 i fodelau trefn 4. Fodd bynnag, mae llai o welliant i drefnau 5 a 6 sy'n perfformio ar lefel debyg.

Yn achos CorCenCC, ar y llaw arall, mae'r darlun yn wahanol iawn. Mae gwelliant cyson o ran cywasgu hyd at fewnbwn o 0.10 Mbeit, ond wedyn mae'r cywasgu'n lleihau'n amlwg. Dydy'r gymhareb cywasgu ddim yn gostwng eto tan tua 0.70 Mbeit. Mae cynnydd amlwg pellach o ran y gymhareb cywasgu yn digwydd ar 1.0 a 7.5 Mbeit. Dengys hyn bod y corpws yn cynnwys testun o wahanol genres o wahanol barthau sy'n gwyro'n sylweddol oddi wrth ei gilydd drwy'r corpws cyfan.

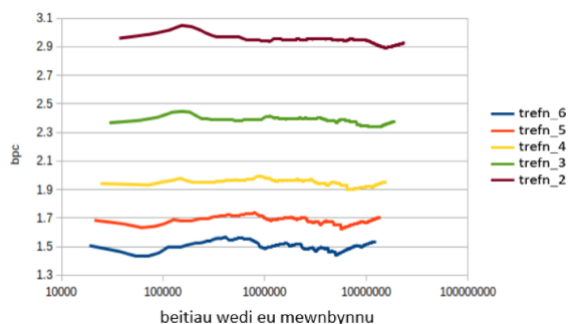
O ganlyniad i hyn, er bod y corpws hwn yn llawer mwy na'r corpws CEG (sydd bron yn 10.613 gwaith yn fwy o ran ei faint), gallai defnyddio'r corpws hwn i hyfforddi modelau iaith ar gyfer NLP ystadegol fod yn llai effeithiol ar gyfer cywasgu testunau prawf gwahanol oherwydd ei natur fwy heterogaidd. Yn lle hyn, gellir cyflawni gwell perfformiad mewn gwahanol dasgau NLP drwy dynnu rhannau unigryw'r corpws a'u defnyddio i greu modelau ar wahân.

3.2 Cywasgu Testun Cymraeg gan Ddefnyddio Modelau Statig

Roedd yr ail gyfres o arbrofion yn edrych ar effeithiolrwydd defnyddio model statig a oedd wedi'i adeiladu o un corpws i wella cywasgiad corpws arall. Yn yr achos hwn, defnyddir offeryn **train** TAWA i breimio model statig, sydd wedyn yn cael ei ddefnyddio i gywasgu'r corpws arall gan ddefnyddio'r offeryn **encode**. Cafodd hyn ei wneud eto ar gyfer trefnau 6 i 2 gan ddefnyddio modelau PPMD. Mae canlyniadau'r arbrofion i'w gweld yn Ffigyrau 4 a 5.



Ffigur 4: Amgodio CorCenCC yn statig gan ddefnyddio CEG ar gyfer hyfforddi.



Ffigur 5: Amgodio CEG yn statig gan ddefnyddio CorCenCC ar gyfer hyfforddi.

Mae'r canlyniadau'n wahanol iawn i'r rhai sydd i'w gweld yn Ffigyrau 2 a 3. Ar yr olwg gyntaf, maen nhw fel petaent yn dangos nad oes rheswm dros ddefnyddio model hyfforddi gan fod yr algorithm cywasgu yn gwneud yn waeth yn y dull statig. Fel y nodwyd uchod, mae'n hysbys bod y model dynamig yn gallu perfformio'n well na'r model statig os yw'n cael digon o ddata hyfforddi, gan fod y model yn un addasol ac yn diweddar ei hun wrth i'r testun gael ei brosesu.

Yn yr achos hwn, mae defnyddio CEG fel y corpws hyfforddi yn syniad gwael wrth amgodio CorCenCC (gweler Ffigur 4) gan nad yw'r model statig penodol yn rhagfynegydd da o'r testun a ragwelir ar ddau brif bwynt, sef y brigau sydyn sy'n digwydd ar oddeutu 200,000 a 20,000,000 beit. Y rheswm mwyaf tebygol am hyn yw'r newidiadau yn y testun o ganlyniad i'r gwahanol genres neu arddulliau sy'n digwydd ar yr adegau hynny.

Mantais defnyddio model statig wedi'i breimio ymlaen llaw yw y gall y broses cywasgu fod yn effeithiol o'r dechrau, ond dim ond os yw'r testun hyfforddi yn gynrychiadol o'r testun sy'n cael ei gywasgu. Er enghraifft, mae'r gymhareb cywasgu yn dechrau ar 1.5 bpc ar gyfer y model trefn 6 wrth amgodio'r corpws CEG (gweler Ffigur 5) yn hytrach nag ar 2.8 bpc yn Ffigur 2. Ar y llaw arall, mae'r gymhareb cywasgu yn dechrau ar 2.7 bpc yn Ffigur 4 sy'n waeth na'r 2.0 bpc yn Ffigur 3. Mae hyn yn dangos nad yw'r testun CEG yn rhagfynegydd da ar gyfer CorCenCC.

Casgliad pwysig o Ffigyrau 2 i 5, fodd bynnag, yw mai modelau trefn 6, ar gyfer amgodio dynamig a statig, sy'n arwain at y canlyniadau gorau ar gyfer y Gymraeg.

3.3 Dosbarthu Testunau Cymraeg

Dangoswyd bod y dull dosbarthu testun yn seiliedig ar gywasgu yn effeithiol mewn nifer o wahanol barthau [1]. Pwrpas yr arbrawf rhagarweiniol yma yw gweld pa mor dda mae'n perfformio wrth ddosbarthu testun Cymraeg. Mae'r dogfennau yn y corpws CEG wedi cael eu dosbarthu â llaw yn ôl genre'r testun, felly gellir defnyddio'r rhain fel y 'gwirionedd sylfaenol' ar gyfer yr arbrawf hwn. Yn benodol, mae'r dogfennau wedi cael eu rhannu'n ddau brif gategori - 'Ffuglen' a 'Ffeithiol', yn ogystal â 21 is-gategori sy'n cynnwys 'Y Wasg - Gwyddonol', 'Academaidd', 'Nofelau' a 'Straeon Byr'.

Ar gyfer yr arbrawf rhagarweiniol hwn, fe wnaethon ni benderfynu peidio â defnyddio'r is-gategorïau gan nad oes gan nifer o'r rhain ddigon o ddogfennau hyfforddi ar gyfer hyfforddi'r modelau PPM. Roeddem wedi defnyddio trawsddilysu 5-plyg er mwyn gwerthuso'n gadarn ar draws y set ddata gyfan. Cafodd ffeiliau prawf eu dosbarthu ym mhob maes gan ddefnyddio modelau PPMD statig a luniwyd gan ddefnyddio data hyfforddi o weddill y set ddata. Dangosir y canlyniadau isod gan ddefnyddio modelau PPMD gwahanol (o drefn 2 i drefn 6). TP, FP, FN a TN yw nifer y gwir bositifau, y positifau ffug, y negatifau ffug a'r gwir negatifau yn y drefn honno wrth ystyried y dasg fel problem dosbarthu ddeuaidd, gyda 'Ffuglen' yn ddosbarth positif a 'Ffeithiol' yn ddosbarth negyddol. Mae cywirdeb yn cael ei gyfrifo gyda $(TP + TN) / (TP + TN + FP + FN)$, adalw gyda $TP / (TP + FN)$, trachywirededd gyda $TP / (TP + FP)$ a'r sgôr F1 gyda $(2 * TP) / (2 * TP + FP + FN)$.

Tabl 2: Canlyniadau'r arbrawf rhagarweiniol

Gorchymyn	TP	FP	FN	TN	Cywirdeb	Adalw	Trachywirededd	Sgôr F1
2	89	43	15	350	0.883	0.856	0.674	0.754
3	89	32	15	361	0.905	0.856	0.736	0.791
4	75	13	29	380	0.915	0.721	0.852	0.781
5	58	6	46	387	0.895	0.558	0.906	0.690
6	40	4	64	389	0.863	0.385	0.909	0.541

Mae'r canlyniadau'n dangos bod y dosbarthwr sy'n seiliedig ar gywasgiad yn perfformio â lefel uchel o gywirdeb (0.915), yn enwedig pan ddefnyddir modelau trefn 4. Mae adalw ar ei uchaf gyda modelau trefn 2 a 3 (0.856), mae trachywirededd ar ei uchaf gyda modelau trefn 6 (0.909) ac mae'r Sgôr F1 ar ei uchaf gyda modelau trefn 3 (0.791). Mae'r canlyniadau'n dangos bod y dull ar sail cywasgu yn effeithiol iawn o ran gwahaniaethu rhwng testun Cymraeg ffuglennol a ffeithiol.

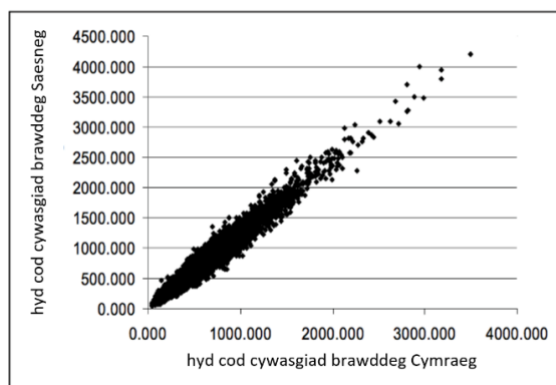
Mae canlyniadau arbrofol eraill ar gyfer gwahanol ieithoedd a pharthau (fel genre a rhywedd) yn dangos bod modelau ar sail cywasgu yn gystadleuol â dulliau dysgu peirianyddol o'r radd flaenaf neu'n rhagori arnynt. Byddwn yn gwneud cymhariaeth arbrofol lawn gyda'r dulliau hyn rywbryd yn y dyfodol.

3.4 Gwirio Aliniad Brawddegau Mewn Testun Cymraeg-Saesneg Cyfochrog gan Ddefnyddio Hyd Cod Cywasgu

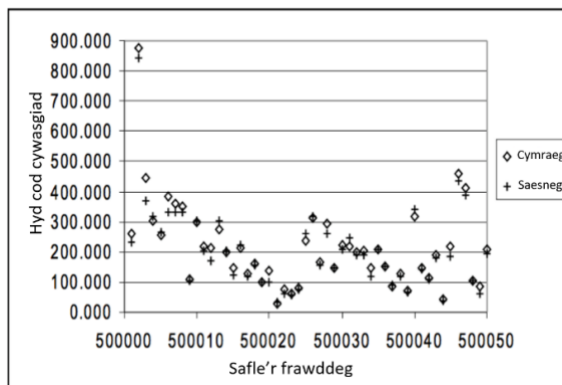
Mae'r adran hon yn adrodd ar arbrofion i wirio aliniad brawddegau'r corpws UAGAT-PNAW. Gwneir hyn gan ddefnyddio hyd cod cywasgu i fesur cynnwys y wybodaeth ym mhob llinyn nod testun, wedi'i amcangyfrif gan ddefnyddio'r offeryn **codlength** ym mhecyn cymorth TAWA. Mae'r syniad y tu ôl i'r dull hwn yn dibynnu ar y rhagosodiad y dylai cywasgu testun sydd wedi'i gydgyfieithu (h.y. brawddegau yn yr achos hwn) fod yn debyg o ran hyd codau [1] gan y bydd y wybodaeth yn y cydgyfieithiadau yn debyg. Gan fod modd defnyddio cywasgu i fesur

cynnwys y wybodaeth, gallwn edrych ar gymhareb hyd cod cywasgu'r pâr testunau sydd wedi'i gyfieithu ar y cyd i benderfynu a yw'r testun wedi'i alinio. Hynny yw, os oes gennych llyn testun brawddeg mewn un iaith, a'i gyfieithiad mewn iaith arall, yna dylai'r gymhareb R o hyd cod cywasgu'r pâr llinynnau testun fod yn agos at 1.0, pan $R = \max(C_W/C_E, C_E/C_W)$ a C_E yw'r hyd cod cywasgu ar gyfer y frawddeg Gymraeg a C_W yw'r hyd cod cywasgu ar gyfer y frawddeg Saesneg.

Mae canlyniadau'r arbrofion rhagarweiniol ar gyfer corpws UAGAT-PNAW i'w gweld yn Ffigyrau 6 i 9. Sylwch fod yr hyd cod cywasgu yn yr achos hwn yn cael ei gyfrifo gan ddefnyddio modelau statig wedi'u neilltuo ymlaen llaw sydd wedi'u hyfforddi ar destun o'r ieithoedd perthnasol. Mae hyn oherwydd bod testunau'r brawddegau yn rhy fyr i fodolau PPM dynamig gynhyrchu canlyniadau y mae modd gwahaniaethu rhyngddynt. Defnyddir y corpws CEG i hyfforddi'r model Cymraeg ar gyfer cywasgu brawddegau Cymraeg, a defnyddir corpws Saesneg LOB i hyfforddi'r model Saesneg ar gyfer cywasgu'r brawddegau Saesneg.



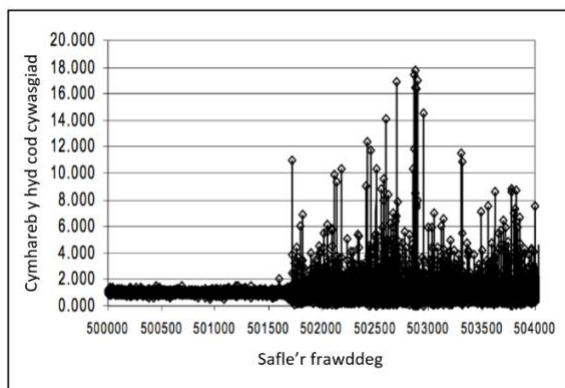
Ffigur 6: Hydoedd cod cywasgu ar gyfer y 10,000 brawddeg gyntaf yng nghorpws cyfochrog UAGT-PNAW.



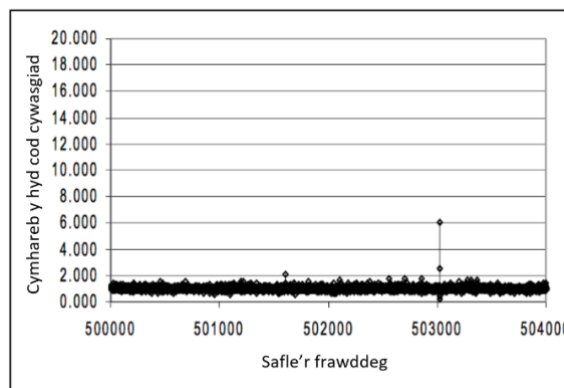
Ffigur 7: Hydoedd cod cywasgu ar gyfer brawddegau yn y corpwscyfochrog UAGT-PNAW.

Mae Ffigur 6 yn graffio cymarebau cywasgu hydoedd cod yn erbyn gwahanol leoliadau brawddegau yn y corpws. Mae'r hydoedd cod ar gyfer y brawddegau Cymraeg a Saesneg wedi'u plotio ar gyfer y lleoliadau brawddegau 500,000 hyd 500,050 yn Ffigur 7. Mae'r ffigyrau'n dangos yn glir bod yr hydoedd cod cywasgu ar gyfer y rhan fwyaf o frawddegau'n cyfateb yn dda iawn ac yn debyg iawn o ran eu gwerth. Mae Ffigur 8 yn plotio cymarebau cywasgu hydoedd cod R ar gyfer yr un lleoliadau brawddegau, sef 500,000 hyd 500,050 ar gyfer y corpws.

Fel y nodwyd, dylai'r cymarebau hyn fod yn agos at 1.0. Mae hyn yn wir hyd at y lleoliad brawddeg 501,719 pan fydd y gymhareb yn neidio'n sydyn i dros 10.0. Ar ôl hyn, mae'n amlwg nad yw'r brawddegau canlynol yn cydfynd â'i gilydd. Unwaith y mae'r camgymeriad hwn wedi'i gywiro, mae'r brawddegau canlynol wedyn yn cael eu cysoni'n gywir ac eithrio'r camgymeriad posibl pellach amlwg tua lleoliad y frawddeg 503,000.



Ffigur 8: Cymarebau hydoedd cod cywasgu ar gyfer brawddegau yn y corpwscyfochrog UAGT-PNAW.



Ffigur 9: Ar ôl cywiro gwall yn safle'r frawddeg 501719.

4 TRAFODAETH A GWAITH YN Y DYFODOL

Mae'r bennod hon wedi ymchwilio i gymhwyso'r pecyn cymorth ar sail cywasgu TAWA i broblemau amrywiol mewn NLP Cymraeg. Defnyddiwyd tri chorpws Cymraeg yn y gwerthusiad arbrofol: y corpws CEG [2]; y corpws CorCenCC [4], [5] sydd wedi ymddangos yn ddiweddar; a'r corpws dwyieithog Cymraeg-Saesneg UAGT-PNAW [6]. Mae'r canlyniadau'n dangos bod modd defnyddio'r adnoddau yn y pecyn cymorth i berfformio tair tasg yn effeithiol: cywasgu; categorio testunau; ac alinio brawddegau ar gyfer testunau cyfochrog. Mae'r gwaith hwn yn rhoi sylfaen bwysig i gymhwyso dulliau NLP sy'n seiliedig ar gywasgu i'r Gymraeg yn ffurfiol.

Mae llawer mwy o gymwysiadau'n bosibl a byddwn yn ymchwilio iddynt mewn gwaith sydd i ddod. Bydd hyn yn cynnwys ymchwilio i ganfod cyfnewid cod yn awtomatig gan ddefnyddio offeryn markup TAWA gan fod gan y corpws CEG enghreifftiau o gyfnewid cod o'r Gymraeg i'r Saesneg yn ogystal ag o'r Gymraeg i'r Ffrangeg. Ar ben hynny, byddai'n ddiddorol diweddar CySill gan ddefnyddio CorCenCC gan ei fod yn darparu set ddata llawer mwy i hyfforddi arni.

Ni ellir tanbriso pwysigrwydd cymryd i ystyriaeth amrywiaeth ieithyddol y Gymraeg. Roedd defnyddio CEG fel model hyfforddi ar gyfer amgodio CorCenCC yn wers dda o ran ansawdd y canlyniadau o'i gymharu â maint ac amrywiaeth y corpws. Fodd bynnag, roedd y canlyniadau a gafwyd o ddefnyddio CorCenCC fel model hyfforddi ar gyfer amgodio'r corpws CEG yn drawiadol.

CYFEIRIADAU

- [1] William J. Teahan. 2018. A Compression-Based Toolkit for Modelling and Processing Natural Language Text. Information, 9, 12 (Tachwedd 2018). <http://doi:10.3390/infoxx010001>.
- [2] John G. Cleary ac Ian H.Witten. 1984. Data compression using adaptive coding and partial string matching. IEEE Transactions on Communications, 32, 4 (Ebrill 1984), 396-402.
- [3] Nick C. Ellis, C. O'Dochartaigh, William Hicks, M. Morgan a N. Laporte. 2001. Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh. [Ar-lein]
- [4] Dawn Knight, Steve Morris, Tess Fitzpatrick, Paul Rayson, Irena Spasić, Enlli-Mon Thomas, Alex Lovell, Jonathan Morris, Jeremy Evas, Mark Stonelake, Laura Arman, Josh Davies, Ignatius Ezeani, Steve Neale, Jennifer Needs, Scott Piao, Mair Rees, Gareth Watkins, Lowri Williams, Vignesh Muralidaran, Bethan Tovey-Walsh, Laurence Anthony, Thomas M. Cobb, Margaret Deuchar, Kevin Donnelly, Michael McCarthy a Kevin Scannell. 2020. CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh Dataset. <http://doi.org/10.17035/d.2020.0119878310>.
- [5] Dawn Knight, Steve Morris, Tess Fitzpatrick, Paul Rayson, Irena Spasić ac Enlli-Mon Thomas. 2020. The National Corpus of Contemporary Welsh: Project Report | Y Corpws Cenedlaethol Cymraeg Cyfoes: Adroddiad y Prosiect. Prifysgol Caerdydd, Cymru. ar Xiv:2010.05542.
- [6] Dafydd Jones ac Andreas Eisele. 2006. Phrase-based statistical machine translation between English and Welsh. Yn Strategies for developing machine translation for minority languages (5th SALT MIL workshop on Minority Languages), LREC-2006. LREC, Genoa, Yr Eidal, 75-78.

Testun-i-Leferydd Ddwyieithog Cymraeg a Saesneg

DEWI BRYN JONES,

Uned Technolegau Iaith, Prifysgol Bangor, Cymru

DR SARAH COOPER

Ysgol Iaith, Diwylliant a'r Celfyddau, Prifysgol Bangor, Cymru

Mae'r bennod hon yn disgrifio gwaith ar ddatblygu system testun-i-leferydd ddwyieithog Cymraeg-Saesneg drwy gyfuno geiriaduron ynganu Cymraeg a Saesneg oedd ar gael eisoes a thrwy wella ansawdd corpws lleferydd bach ddwyieithog. Ehangwyd y geiriadur ynganu ddwyieithog newydd gyda geiriau newydd drwy gymorth modelau graffem-i-ffonem newydd a gafodd eu hyfforddi a'u gwerthuso am eu cywirdeb. Adeiladwyd y llais gyda modiwl synthesis dethol unedau MaryTTS sydd, yn ôl gwrandio rhagbrofol yn unig, yn cynhyrchu lleferydd Cymraeg a Saesneg dealladwy. Fodd bynnag, amlygir rhai cyfyngiadau ynghyd â thrafodaeth am waith pellach.

1 CYFLWYNIAD

Mae systemau testun-i-leferydd yn galluogi cyfrifiaduron i gynhyrchu fersiynau sain o destun mewn iaith benodol o fewn ystod eang o raglenni fel darlennydd tudalennau gwe, rhyngwyneb defnyddwyr sydd wedi'i wella gan leferydd, cynorthwydd personol digidol a systemau cyhoeddi wedi'u hawtomeiddio [1], [2].

Ers rhai blynedd, mae amrywiaeth o systemau yn cynhyrchu sain o destunau Cymraeg. Crëwyd testun-i-leferydd Cymraeg ar sail deuffonau gyda'r system synthesis lleferydd Festival [3] yn 2006 o ganlyniad i gydweithrediad rhwng Prifysgol Bangor a phrifysgolion Dulyn yn ystod project Interreg UE o'r enw WISPR (Welsh and Irish Speech Processing Resources). Datblygodd WISPR yn ogystal nifer o adnoddau lleferydd at ddibenion ymchwil bellach technoleg lleferydd Cymraeg a Gwyddeleg [4]. Yn 2009 cyhoeddwyd cynnyrch testun-i-leferydd Cymraeg masnachol gan y cwmni Ivona gydag ansawdd synthesis llawer mwy naturiol. Mae'r ddarpariaeth bellach ar gael o fewn gwasanaeth lleferydd Amazon Polly. Yn ddiweddarach yn 2016 datblygwyd llais cod agored oedd hefyd yn swinio'n fwy naturiol gyda'r system MaryTTS [5] ar gyfer defnydd cynorthwydd digidol personol Cymraeg - Macsen [6] a rhaglen bancio lleferydd o'r enw Lleisiwr¹.

Fodd bynnag, mae defnyddwyr testun-i-leferydd Cymraeg yn byw ac yn cyfathrebu mewn amgylchoedd ddwyieithog sydd angen cynhyrchu sain o destunau Saesneg hefyd. Nid yw defnyddio a newid rhwng dwy system testun-i-leferydd uniaith ar wahân yn ymarferol nac yn ddigonol ar gyfer ynganu geiriau benthyg, rhai enwau priodol a hefyd ar gyfer achosion o gyfnawid cod. O ganlyniad, byddai cael un system testun-i-leferydd, sy'n gallu cynhyrchu sain llais sy'n swinio'n naturiol ar gyfer testunau o'r ddwy iaith, yn beth da iawn. Gall pwrcasu a chynnal un system testun-i-leferydd, yn hytrach na dwy, bod yn fanteisiol i sefydliadau sydd angen systemau cyhoeddiadau

¹<https://lleisiwr.cymru>

cyhoeddus a hefyd i ddefnyddwyr y rhaglen Lleisiwr fyddai'n elwa o hwylustod proses bancio a hyfforddi mwy effeithlon.

2 DADANSODDIAD ADNODDAU

Mae'r dull synthesis lleferydd a ddewisir fel ateb i'r broblem yn dylanwadu ar ba adnoddau sydd eu hangen. Penderfynwyd gwireddu synthesis testun-i-leferydd dwyieithog cychwynnol gyda'r adnoddau agored o'r gwaith a wnaed eisoes gyda Festival a MaryTTS. Defnyddiwyd modiwl synthesis lleferydd dethol unedau MaryTTS, sydd angen dadansoddiad seinegol o iaith, sef geiriadur ynganu, a recordiadau o ansawdd uchel gan actor yn darllen sgript recordio wedi'i gynllunio'n arbennig.

2.1 Geiriaduron Ynganu

Gan ein bod yn dymuno cael system testun-i-leferydd ddwyieithog roedd angen i ni greu geiriadur ynganu dwyieithog oedd yn cynnwys niferoedd mawr o eiriau o'r ddwy iaith wedi'u disgrifio yn seinegol gyda chyfres o ffonemau. Roedd hyn angen un set o ffonemau er mwyn cynrychioli pob un sain bosibl o'r ddwy iaith.

Penderfynwyd gwireddu geiriadur dwyieithog drwy gyfuno'r geiriau a ffonemau o eiriaduron agored oedd eisoes ar gael ar gyfer y Gymraeg a'r Saesneg ar sail efelychu dulliau ffurfio uwchset ffonemau gan eraill. [7]. Dewiswyd geiriadur ynganu llais Festival Cymraeg a'r geiriadur Saesneg Americanaidd CMU Pronouncing Dictionary².

2.1.1 Geiriadur Ynganu Llais Cymraeg Festival

Mae'r geiriadur ynganu o fewn y llais Festival Cymraeg yn cynnwys trawsgrifiadau seinegol i bron 39000 o eiriau gyda set ffonemau sy'n cynnwys 14 llafariad byr a hir, 35 cytsain a 14 deusain.

Gellir gweld trawsgrifiad enghreifftiol yn nhabl 1. Cynrychiolir ffonemau fel nodau ASCII yn ôl y Machine-Readable Phonetic Alphabet for Welsh (MRPAW) fel y'i cynlluniwyd yn Williams [8]. Mae cromfachau yn amgau sillafau. Nodir pwyslais gyda'r rhif 1 tra bo 0 yn cyfleu bod dim pwyslais.

Mae'r geiriadur wedi tyfu mewn maint yn araf dros nifer o flynyddoedd gyda geiriau wedi'u hychwanegu a'u trawsgrifio'n ffonetig â llaw. Roedd dadansoddi'r geiriadur drwy gymharu gyda geirfaoedd gwirydd sillafu Hunspell³ Cymraeg a Saesneg yn datgelu ei fod yn cynnwys tua 26000 o eiriau Cymraeg, 12000 o eiriau Saesneg, 2600 o'r ddwy iaith tra bo tua 3500 ddim yn cael eu hadnabod fel geiriau Gymraeg na Saesneg. Mae'r rhestr o eiriau Cymraeg yn cynnwys lemata, treigladau a ffurfdroadau. Nid yw pob ffurfdroad o lema yn bresennol. Mae'r geiriadur yn cynnwys maes ychwanegol ar gyfer tagio geiriau gydag un o 61 elfen meta-data bosibl, e.e. rhan ymadrodd. Ond nid yw'r maes ychwanegol yn cofnodi iaith y gair. Mae'r tag 'foreign' yn nodi geiriau benthyg o'r Saesneg neu sillafiad yn ôl orgraff Gymraeg o air Saesneg gwreiddiol e.e. 'pyjamas', 'telefision'. Fel arall, mae geiriau Saesneg yn defnyddio'r un rhannau ymadrodd â geiriau Cymraeg.

²The CMU Pronouncing Dictionary, ar gael o <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

³Hunspell Cymraeg, ar gael o <http://techiaith.cymru/data/geiriadurol/hunspell-cymraeg/>

Mae bodolaeth a nifer y geiriau Saesneg yn awgrymu y bu ymdrech i alluogi lleisiau Festival Cymraeg i ynganu o leiaf set gyfyngedig o eiriau Saesneg o'r set ffonemau Cymraeg – er bod un ffonem ychwanegol wedi'i chynnwys ar gyfer nodi ffrithiolyn ôl-orfannol lleisiol (zh /3/) mewn geiriau fel 'explosion', 'measure' a 'television'.

Nid ydym yn gwybod beth oedd sail ychwanegu a thrawsgrifio'n seinegol y geiriau Saesneg hyn. Mae'r nifer yn gymharol fach i'w gymharu ag anghenion geiriadur ynganu llawn. Ni allwn ddisgwyl y bydd hyn yn ddigonol ar gyfer ynganu amrywiaeth eang o destunau Saesneg o fewn lleisiau testun-i-leferydd dwyieithog.

2.1.2 Geiriadur Ynganu'r CMU

Mae geiriadur ynganu'r CMU (CMUDict) yn eiriadur ynganu cod agored poblogaidd a grëwyd gan y grŵp ymchwil lleferydd ym Mhrifysgol Carnegie Mellon. Mae'n cynnwys trawsgrifiadau seinegol ynganiadau Gogledd America o eiriau Saesneg. Gellir gweld trawsgrifiad enghreifftiol yn nhabl 1.

Cynrychiolir ffonemau fel nodau ASCII gan ddefnyddio codau ARPABET⁴. Defnyddir maes ychwanegol, sy'n cynnwys rhif yn unig, ar gyfer nodi ynganiad amgen i air. Nid yw CMUDict yn nodi unrhyw wybodaeth sillafau. Nodir pwyslais drwy atodi 1 (prif bwyslais) neu 2 (ail bwyslais) i'r ffonem berthnasol.

Yn y fersiwn ddiweddaraf - 0.7b (cyhoeddwyd 2014) - mae CMUDict yn cynnwys tua 138,000 o gofnodion.

Tabl 1 – Cymharu enghreifftiau o eiriaduron ynganu Cymraeg a Saesneg

Gair	Geiriadur Ynganu	Disgrifiad Seinegol
testun-i-leferydd	Festival (Cymraeg)	adj (((t e s) 0) ((t y n) 0) ((i i l) 0) ((e v) 0) ((e r) 1) ((i dh) 0))
Text-to-Speech	CMUDict (Saesneg)	T EH1 K S T . T UW1 . S P IY1 CH .

Mae CMUDict wedi cael ei ddefnyddio fel corpws hyfforddi ar gyfer adeiladu modelau graffem-i-ffonem (G2P) sy'n gallu cynhyrchu disgrifiadau seinegol ar gyfer geiriau newydd nad ydynt yn y geiriadur.

2.2 Data Lleferydd Corpws WISPR

Cynhyrchodd project WISPR (Williams et. al., 2006) recordiadau stiwdio lleferydd o ansawdd da a fwriadwyd fel adnodd hyfforddi synthesis lleferydd dethol unedau Cymraeg, er bod prif ffocws y project ar wireddu synthesis lleferydd deuffonau. Mae bellach ar gael o'r Porth Technolegau Iaith Cenedlaethol. Y gwaith hwn yw'r ymgais gyntaf o bwys i'w ddefnyddio ar gyfer hyfforddi testun-i-leferydd.

Nid oes lawer o ddogfennaeth na metadata yn y corpws. Mae gan y corpws recordiadau o 265 frawddegau, rhai hir iawn, o'r Beibl Cymraeg yn ogystal â 351 recordiad o frawddegau o draethawd hir myfyriwr israddedig. Mae rhai brawddegau yn gyfan gwbl neu'n rhannol yn Saesneg. Mae cyfanswm o 3 awr ac 30 munud o recordiadau gyda 20 munud yn cynnwys recordiadau o frawddegau Saesneg cyfan. Mae'r talent lais yn wrywaidd gydag acen gogledd Cymru.

⁴ARPABET - <https://en.wikipedia.org/wiki/ARPABET>

⁵Corpws WISPR - <http://techiath.cymru/data/corpora/wispr/>

3 GWIREDDU TESTUN-I-LEFERYDD DDWYIEITHOG

3.1 Creu Uwchset Ffonemau

Cyfieithwyd pob ffonem o'r ddau eiriadur i'w ffonemau cyfatebol safonol o'r IPA⁶.

O ganlyniad, roeddem wedi sylweddoli bod ffonemau cytseiniaid yn MRPAW ac ARPABET yn cyfateb yn dda iawn ond bod gwahaniaethau mwy sylweddol rhwng ffonemau llafariad a deuseiniaid ynddynt. Gan fod MRPAW yn cynnwys mwy o amrywiaeth o ffonemau roedd modd i ni gysylltu ffonemau unigryw CMUDict i ffonemau a ddyfarnwyd yn ddigon tebyg yn MRPAW yn hytrach na ffurfio uwchset syml. Roedd MRPAW yn annigonol ar gyfer nodi dim ond un ffonem unigryw CMUDict, sef y llafariad rhotig (/ɜ/) fel y glywir yn y geiriau Saesneg 'hurt', 'mirror' a 'water'.

Cynlluniwyd MRPAW yn wreiddiol ar gyfer trawsgrifio'r Gymraeg gydag acen de Cymru. O ganlyniad roedd yn cynnwys tair cytsain wedi'u gwefusoli (/lw/, /nw/ a /rw/). Barnwyd fod y rhain yn ddibwys ac yn amherthnasol i acenion eraill ac ar gyfer testun-i-leferydd Cymraeg yn gyffredinol. Mae astudiaethau eraill ar ffonoleg y Gymraeg [9]; [10]; [11] hefyd yn hepgor cynnwys cytseiniaid wedi'u gwefusoli.

Gweler yr atodiad i weld pob set ffonemau terfynol a'u cysylltiad i ffonemau Festival Cymraeg a CMUDict.

3.2 Creu Geiriadur Ynganu Dwyieithog Cytbwys

Mae'r ddau eiriadur ynganu yn wahanol iawn o ran maint gyda CMUDict yn cynnwys bron i 100k o eiriau yn fwy na geiriadur Festival Cymraeg. Er mwyn osgoi tuedd tuag at un iaith o fewn unrhyw system testun-i-leferydd ddwyieithog a pheri ynganu geiriau yn anghywir ceisiwyd cael cytbwysedd o'r un nifer o eiriau o'r ddwy iaith. Cyflawnwyd hyn drwy gynyddu'r nifer o eiriau Cymraeg a lleihau'r nifer o eiriau Saesneg i gael casgliad o eiriau mwyaf cyffredin.

Yn anffodus, nid oes rhestrau parod o amllder geiriau Saesneg o fewn corpora cyfoes enfawr ar gael yn rhwydd a di-gost gyda thrwydded ganiataol er mwyn ffurfio rhestr geiriau mwy cyffredin, er bod rhestrau ar gael i'w prynu gyda chyfyngiadau aildosbarthu anaddas [12]. Aethom ati i gynhyrchu rhestr amllder geiriau Saesneg ein hunain gydag adnoddau agored a hwylus llyfrgell NLTK [13] a chorpws Brown [14]. Defnyddiwyd hefyd rhestr 500 gair ysgrifenedig Saesneg amlaf eu defnydd yn Gymraeg [15]. O ganlyniad, mae'r geiriadur ynganu ddwyieithog cytbwys yn cynnwys hyd at 20000 air Saesneg ac hyd at 30000 gair Cymraeg.

Roedd angen datrysiad i gydbwyso gwybodaeth sillafau geiriau. Mae geiriadur ynganu Festival Cymraeg yn dynodi sillafau o fewn cromfachau, ond nid yw CMUDict yn cynnwys unrhyw wybodaeth o gwbl. Gwnaethom ganfod a defnyddio canlyniadau gwaith [16] ar ddefnyddio dulliau dilyniannu seinlawnder (*sonority sequencing methods*) er mwyn cynhyrchu sillafau ar gyfer ein detholiad geiriau o eiriadur CMUDict.

Gellir gweld cofnodion enghreifftiol isod yn ffigur 1. Mae'r fformat yn darparu pedair colofn sy'n hawdd i'w dosrannu yn gyfrifiadurol:

1. Y gair neu'r graffemau
2. Metadata, wedi'u hamgáu gan gromfachau, ar gyfer nodi iaith, rhan ymadrodd neu ffynhonnell.
3. Ffonemau ar ffurf nodau ASCII (colofn MRPAW yn nhabl yr atodiad)
4. Ffonemau ar ffurf nodau o'r wyddor IPA, wedi'u hamgáu gan flaen-slaes

⁶The International Phonetic Association - <https://www.internationalphoneticassociation.org/>

Gwahanir ffonemau gan fwllch, sillafau gan gysylltnod, gyda chollnod yn dangos lleoliad pwyslais.

syml 's @ - m y l /'sə.mil/
 symlaf 's @ m - l a v /'sə.m.lav/
 symleiddio s @ m - 'l ei dh - j o /sə.m.'ləið.jo/
 symlirwydd 's @ m l - r uy dh /'səml.ruið/
 symol 's @ - m o l /'sə.mol/
 symposiwm s @ m - 'p oo - s iu m /sə.m.'pɔ:.siu.m/
 symptomau s @ m p - 't o - m ay /səmp.'tɔ.mai/

 symphonies (en.cmu) s 'i m - f @ - n ii z /s'ɪm.fə.ni:z/
 symphony (en.cmu) s 'i m - f @ - n ii /s'ɪm.fə.ni:/
 symposium (en.cmu) s i m - p 'ou - z ii - @ m /sɪm.p'əu.zi:.əm/
 symptom (en.cmu) s 'i m p - t @ m /s'ɪmp.təm/
 symptomatic (en.cmu) s 'i m p - t @ - m 'a - t i k /s'ɪmp.tə.m'a.tɪk/
 symptoms (en.cmu) s 'i m p - t @ m z /s'ɪmp.təmz/
 synagogue (en.cmu) s 'i - n @ - g 'oo g /s'ɪ.nə.g'ɔ:g/

Ffigur1 - Enghreifftiau o gofnodion o'r geiriadur ynganu dwyieithog

3.3 Hyfforddi Modelau G2P

Roedd angen ychwanegu geiriau Cymraeg i'r geiriadur ynganu ac felly hyfforddwyd amrywiaeth o fodelau graffem-i-ffonem (G2P) er mwyn osgoi gorfod gwneud trawsgrifiad seingol o bob un â llaw. Penderfynwyd yn y gwaith hwn i arbrofi gyda gwahanol gasgliadau o ddata hyfforddi ac nid i werthuso gwahanol ddulliau o fodelu G2P. Defnyddiwyd yn unig Phonetisaurus [17], sy'n llyfrgell o led safonol a chod agored safonol.

Gwelir canlyniadau ein harbrofion yn nhabl 2 isod. Defnyddiwyd dull profi trawsddilysu k-fold (*k-fold cross validation*) er mwyn rhannu'r maint bychan o ddata i'r setiau hyfforddi a phrofi gorau ar gyfer mesur gallu'r model i wneud trawsgrifiad seingol cywir o'r geiriau newydd. Mesurwyd cywirdeb gyda'r metrig cyfradd geiriau gwallus (*word error rate*, WER) sef y nifer o drawsgrifiadau seingol anghywir wedi'u rhannu â chyfanswm y nifer o eiriau yn y set profi.

Gwelwn o'r canlyniadau hyn fod modelau G2P iaith benodol yn trawsgrifio seing yn well na'r modelau dwyieithog. O ganlyniad defnyddiwyd y model Festival Cymraeg i hwyluso ychwanegu geiriau Cymraeg i'r geiriadur ynganu a'r model CMUDict (geiriadur cyfan) ar gyfer geiriau Saesneg newydd.

Tabl 2 – Canlyniadau modelau G2P

Set Hyfforddi	Maint	WER
Festival Cymraeg	28764	25.76%
CMUDict (set geiriau amlaf eu defnydd)	20622	35.34%
Festival Cymraeg + CMUDict (set geiriau amlaf eu defnydd)	49386	34.30%
CMUDict (geiriadur cyfan)	119305	32.65%

3.4 Cywiriadau Corpws Data Lleferydd WISPR

Roedd ein dadansoddiad o gorpws lleferydd WISPR wedi canfod rhai problemau a gwallau a all beri tanseilio ei ddefnyddioldeb ar gyfer hyfforddi system testun-i-leferydd.

Aethpwyd ati i gywiro gwallau sillafu yn ogystal â chael gwared ar atalnodi gormodol. Roedd cryn dipyn o frawddegau'r traethawd hir israddedig yn cynnwys rhifau, fel blwyddyn, ar ffurf digidau. Troswyd y rhain i gyd i'r ffurf geiriau fel y'u hyngener gan y talent lais yn y recordiadau.

Yn olaf, casglwyd geirfa Cymraeg a Saesneg yr holl frawddegau o'r corpws ar gyfer eu hychwanegu, gyda chymorth modelau G2P gorau, i'r geiriadur ynganu ddwyieithog.

3.5 Hyfforddi MaryTTS

Mae MaryTTS [5] yn llyfrgell synthesis testun-i-leferydd poblogaidd ymysg ymchwilwyr academiaidd yn ogystal â datblygwyr meddalwedd cod agored a masnachol. Gan ein bod eisoes wedi defnyddio MaryTTS fel sail i ddatblygu testun-i-leferydd uniaith Gymraeg, roedd adnoddau a chydrannau ar gael i'w ddefnyddio eto ar gyfer datblygu'r synthesis dwyieithog Cymraeg-Saesneg cyntaf.

Ehangwyd y ffeil XML (`allophones.cy.xml`) a oedd eisoes ar gael yn narpariaeth MaryTTS yn Cymraeg i fanylu ar y ffonemau o fewn y geiriadur ynganu ddwyieithog newydd. Ychwanegwyd gwybodaeth bellach am nodweddion sain yr holl ffonemau fel uchder, math o gytsain, lleoliad mynegiant a lleisio. Gweler darnau enghreifftiol yn atodiad y bennod hon a'r ffeil gyfan at GitHub⁷.

Gyda'r geiriadur ynganu dwyieithog newydd a data o gorpws diwygiedig WISPR roedd modd hyfforddi MaryTTS i greu'r testun-i-leferydd dwyieithog Cymraeg-Saesneg.

4 GWERTHUSO'R LLAIS

Llwyddwyd i wneud gwrandawriad rhagbrofol yn unig o'r llais Cymraeg-Saesneg dwyieithog newydd yn y gwaith hwn. Ond sylwyd:

- bod y lleferydd Cymraeg a Saesneg a gynhyrchir gan MaryTTS yn ddealladwy
- bod modd cynhyrchu lleferydd dealladwy o frawddegau hir a heriol erthyglau newyddion a llenyddiaeth Gymraeg a Saesneg
- mai anaml iawn roedd ffonemau neu eiriau coll yn y lleferydd
- bod ynganiad enghreifftiau o gymysgu ieithoedd o fewn testun, fel enwau priodol a chyfnewid cod, ar y cyfan yn gywir
- bod ynganiad homograffau trawsieithol, fel 'wall' a 'plant' yn aml yn yr iaith anghywir.

Mae modd arbrofi gyda'r llais ar dudalen testun-i-leferydd gwefan y Porth Technolegau Iaith⁸.

5 CASGLIAD

Mae'r bennod hon wedi adrodd sut y crëwyd testun-i-leferydd dwyieithog Cymraeg-Saesneg drwy welliannau i adnoddau testun-i-leferydd cod agored oedd yn bodoli eisoes a'r llyfrgell MaryTTS.

Roedd y llais dwyieithog newydd, yn ôl ymarferiad gwrandao rhagbrofol, yn ddigon dealladwy a manteisiol. Mae felly wedi'i ei fabwysiadu yn barod o fewn cynnyrch a gwasanaethau'r Uned Technolegau Iaith fel Lleisiwr a Macsen. Fodd bynnag, mae angen gwerthusiad manwl ohono gyda chymorth gwrandawyr dynol er mwyn canfod mesuriad 'Mean Opinion Score' (MOS) a'i gymharu gyda lleisiau uniaith.

⁷<https://github.com/techiaith/marytts/blob/marytts-lang-cy/marytts-languages/marytts-lang-cy/lib/modules/cy/lexicon/allophones.cy.xml>

⁸Demo o'r llais ddwyieithog ar wefan Porth Technolegau Iaith - <http://techiaith.cymru/leferydd/testun-i-leferydd/>

Datblygwyd geiriadur ynganu dwyieithog newydd yn ogystal â'r modelau G2P gorau posibl er mwyn hwyluso ychwanegu geiriau newydd. Canfu dadansoddiad syml fod tua 2600 o homograffau trawsieithol yn bodoli rhwng y ddwy iaith a bod y llais dwyieithog newydd yn aml yn eu hynganu'n anghywir. Mae angen gwaith pellach ar gadarnhau iaith geiriau o fewn testunau.

Cafodd ansawdd corpws lleferydd WISPR ei wella a'i wneud yn fwy defnyddiol i bawb.

Bwbriedir creu rhagor o leisiau testun-i-leferydd dwyieithog Cymraeg-Saesneg gyda chasgliadau mwy o recordiadau gan 4 talent llais newydd (2 gwryw gydag acenion gogleddol a de, 2 benyw gydag acenion gogleddol a de) gyda'r nod o gynyddu'r amrywiaeth o leisiau a gwella ar ansawdd a naturioldeb lleisiau testun-i-leferydd. Cynlluniwyd y sgriptiau recordio ar eu cyfer gyda chymorth geiriadur ynganu dwyieithog newydd y gwaith hwn.

Cyfeiriadau

- [1] Daniel Jurafsky a James H. Martin. 2014. *Speech and language processing* (2nd. ed., Pearson new internat. ed.). Pearson Education International. Harlow.
- [2] Dewi Bryn Jones, Delyth Prys, Myfyr Prys, a Gruffudd Prys. 2019. *Llawlyfr technolegau iaith*. Coleg Cymraeg Cenedlaethol. Bangor, Cymru.
- [3] Alan W. Black, a Kevin A. Lenzo. 2000. Building voices in the Festival speech synthesis system. *Adalwyd Medi* 5, 2020 o <http://www.festvox.org/festvox-1.1/>
- [4] Briony Williams, Rhys J. Jones, ac Ivan Uemlianin. 2006. Tools and resources for speech synthesis arising from a Welsh TTS project. *Yn Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. European Language Resources Association, Genoa, Yr Eidal, 2574-2577.
- [5] Mark Schröder a Jürgen Trouvain. 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *Yn International Journal of Speech Technology* 6, 4 (Hydref 2003), 365-377. <https://doi.org/10.1023/A:1025708916924>
- [6] Dewi Bryn Jones. 2020. *Macsen: A Voice Assistant for Speakers of a Lesser Resourced Language*. *Yn Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. European Language Resources Association, Marseille, Ffrainc, 194-201.
- [7] Slavcho Chungurski, Ivan Kraljevski, Igor Stojanovic a Blerta Prevalla. 2010. Speech Synthesis of Dissimilar Languages Using their Phonetic Superset. *Yn Proceedings of 8th Conference, Digital Speech and Image Processing*. Iriski Venac, Serbia, A 4.8.
- [8] Briony Williams. 1994. Welsh letter-to-sound rules: Rewrite rules and two-level rules compared. *Computer Speech and Language*, vol. 8 (Gorffennaf 1994), 261-277.
- [9] Robert Mayr a Hannah Davies. 2011. A cross-dialectal acoustic study of the monophthongs and diphthongs of Welsh. *Journal of the International Phonetic Association* (Ebrill 2011), 41 1 - 25.
- [10] Glyn E. Jones. 1984. The distinctive vowels and consonants of Welsh. *Yn Welsh phonology: Selected readings*. Gwasg Prifysgol Cymru, Caerdydd, 40-64.
- [11] Sarah Cooper a Laura Arman (editors). 2020. *Cyflwyniad i ieithyddiaeth*, Coleg Cymraeg Cenedlaethol. Y Coleg Cymraeg Cenedlaethol, Caerfyrddin. ePDF 978-1-911528-19-7
- [12] Mark Davies. 2011. Most frequent 100,000 word forms in English (based on data from the COCA corpus). *Cyrchwyd Medi* 5, 2020 o <https://www.wordfrequency.info/>.
- [13] Steven Bird, Edward Loper ac Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.
- [14] Winthrop N. Francis, ac Henry Kučera. (1964, 1971, 1979) *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers* (Brown). Brown University. Providence, Rhode Island.
- [15] Delyth Prys a Dewi Bryn Jones. 2019. Most common wordforms in Welsh and the most common English words used in Welsh. *Cyrchwyd Medi* 5, 2020 o <http://techiaith.cymru/data/lexicons/wordlists-of-the-most-common-wordforms-in-welsh-and-the-most-common-english-words-used-in-welsh/?lang=en>
- [16] Susan Bartlett, Grzegorz Kondrak a Colin Cherry. 2009. On the Syllabification of Phonemes. *Yn Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*. Association for Computational Linguistics, Boulder, Colorado, USA, 308-316.
- [17] Josef Novak, Nobuaki Minematsu a Keikichi Hirose. 2015. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering* 22, 6 (Tachwedd 2016), 907-938. DOI: 10.1017/S1351324915000315.

ATODIAD

Uwchset Ffonemau Geiriadur Ynganu Dwyieithog Cymraeg-Saesneg

Tabl 1: Llafariaid / Vowels

MRPAW	IPA	Disgrifiad	Enghraifft	CMU
a	a	Llafariad isel agored blaen Open front unrounded vowel		AE /æ/ (at)
aa	a:	Llafariad isel agored blaen hir Open front unrounded vowel long	sâl, tad	
e	ɛ	Llafariad canolog blaen Open-mid front unrounded vowel		EH
ee	e:	Llafariad canolog blaen hir Open-mid front unrounded vowel long	mêl	
i	ɪ	Llafariad uchel caeedig blaen Near-close near-front unrounded vowel		IH
ii	i:	Llafariad uchel caeedig blaen hir Near-close near-front unrounded vowel long	ci	IY /i/ (eat)
o	ɔ	Llafariad canolog cefn Open-mid back rounded vowel		AA /ā/ (odd)
oo	ɔ:	Llafariad canolog cefn hir Open-mid back rounded vowel long	tôn	AO /ɔ/ (ought)
u	ʊ	Llafariad uchel caeedig cefn Close back rounded vowel	cwm	UH
uu	u:	Llafariad uchel caeedig cefn hir Close back rounded vowel long	cŵn, stŵr	UW /u/ (two)
@	ə	Llafariad canolog canol (shwa) Mid central vowel (schwa)		AH /e/ (hut)
@@	ɜ:	Llafariad canolog blaen hir Open-mid central unrounded vowel long	word	
yy	ɪ:	Llafariad uchel caeedig canol hir Near-close back rounded vowel long	sych, tŷ	
@r	ər	Llafariaid rhotig (en_US) Rhotic vowel (en_US)	mirror	ER /ɜ/ (hurt)

Tabl 2: Cytseiniau / *Consonants*

MRPAW	IPA	Disgrifiad	Enghraifft	CMU
b	b	Ffrwydrolyn dwywefusol lleisiol		B
		Voiced plosive bilabial		
k	k	Ffrwydrolyn felar di-lais		K
		Voiceless velar plosive		
x	x	Ffrithiolyn wfwlar di-lais		
		Voiceless uvular fricative		
d	d	Ffrwydrolyn gorfannol lleisiol		D
		Voiced alveolar plosive		
dh	ð	Ffrithiolyn deintiol lleisiol		DH
		Voiced dental fricative		
v	v	Ffrithiolyn gwefus-ddeintiol lleisiol		V
		Voiced labiodental fricative		
f	f	Ffrithiolyn gwefus-ddeintiol ddi-lais		F
		Voiceless labiodental fricative		
g	g	Ffrwydrolyn felar lleisiol		G
		Voiced velar plosive		
hh	h	Ffrithiolyn glotal di-lais		HH
		Voiceless glottal fricative		
jh	dʒ	Affrithiolyn ôl-orfannol lleisiol	Jones, John	JH
		Voiced postalveolar affricate		
l	l	Dynesolyn ochrol gorfannol lleisiol		L
		Voiced alveolar lateral approximant		
lh	ɫ	Ffrithiolyn ochrol gorfannol di-lais		
		Voiceless alveolar lateral fricative		
m	m	Trwynolyn dwywefusol lleisiol		M
		Voiced bilabial nasal		
mh	ɱ	Trwynolyn dwywefusol di-lais		
		Voiceless bilabial nasal		
n	n	Trwynolyn gorfannol lleisiol		N
		Voiced alveolar nasal		
nh	ɳ	Trwynolyn gorfannol di-lais		
		Voiceless alveolar nasal		
ng	ŋ	Trwynolyn felar lleisiol		NG
		Voiced velar nasal		
ngh	ɲ	Trwynolyn felar di-lais		
		Voiceless velar nasal		
p	p	Ffrwydrolyn dwywefusol di-lais		P
		Voiceless plosive bilabial		
r	r	Tril gorfannol lleisiol		R /ɹ/ (read)
		Voiced alveolar trill		
rh	ʀ	Tril gorfannol di-lais		
		Voiceless alveolar trill		
s	s	Ffrithiolyn gorfannol di-lais		S
		Voiceless alveolar fricative		
sh	ʃ	Ffrithiolyn ôl-orfannol di-lais		SH
		Voiceless postalveolar fricative		
t	t	Ffrwydrolyn gorfannol di-lais		T
		Voiceless alveolar plosive		

MRPAW	IPA	Disgrifiad	Enghraifft	CMU
th	θ	Ffrithiolyn deintiol di-lais Voiceless dental fricative		TH
ch	tʃ	Affrithiolyn ôl-orfannol di-lais Voiceless postalveolar affricate	cwtsh	CH
z	z	Ffrithiolyn gorfannol lleisol Voiced alveolar sibilant	crazy	Z
zh	ʒ	Ffrithiolyn ôl-orfannol lleisiol Voiced palato-alveolar fricative	explosion	ZH
j	j	Dynesolyn taflodol lleisiol Voiced palatal approximant	wincio	Y
w	w	Dynesolyn felar lleisiol Voiced labial-velar approximant	swigen	W

Tabl 3: Deuseiniaid / Diphthongs

MRPAW	IPA	Disgrifiad	Enghraifft	CMU
ai	aɪ	Deusein cau blaen Front closing diphthong	llais, tai	AY
ei	əɪ	Deusein cau blaen Front closing diphthong	tei, neidio	EY /eɪ/ (ate)
eu	ɛʊ	Deusein cau blaen Front closing diphthong	mewn, llew	
oi	ɔɪ	Deusein cau blaen Front closing diphthong	cnoi	OY
ay	aɪ	Deusain cau canolog Central closing diphthong	haul, cau	
aay	aɪ	Deusain cau canolog Central closing diphthong	cae	
ey	əɪ	Deusain cau canolog Central closing diphthong	creu, neu	
oy	ɔɪ	Deusain cau canolog Central closing diphthong	troed, coed	
uy	ʊɪ	Deusain cau canolog Central closing diphthong	dwyn, mwy	
iu	ɪʊ	Deusain cau cefn Back closing diphthong	brîw	
ou	əʊ	Deusain cau cefn Back closing diphthong	trowsus	OW /ou/ (oat)
yu	ɪʊ	Deusain cau cefn Back closing diphthong	clyw Duw	
au	aʊ	Deusain cau cefn Back closing diphthong	pawb, llaw	AW


```

<allophones name="bangor-mrpaw"
  xml:lang="cy"
  features="vheight vfront vrnd ctype cplace cvox">

  <silence ph="_" />

  <!-- ##### -->
  <!-- Monophthongs / Unseiniaid -->
  <!-- ##### -->

  <!-- open front unrounded vowel /a/ -->
  <vowel ph="a" vc="+" vln="0" vheight="3" vfront="1" vrnd="-" />

  <!-- open front unrounded vowel (long) /a:/ 'sâl tad' -->
  <vowel ph="aa" vc="+" vln="1" vheight="3" vfront="1" vrnd="-" />

  <!-- open mid-front unrounded vowel /e/ -->
  <vowel ph="e" vc="+" vln="0" vheight="2" vfront="1" vrnd="-" />

  <!-- ##### -->
  <!-- Diphthongs / Deuseiniaid -->
  <!-- ##### -->

  <!-- Front closing -->

  <!-- /aɪ/ 'lâis, tai' -->
  <vowel ph="ai" vc="+" vln="d" vheight="1" vfront="1" vrnd="-" />

  <!-- /eɪ/ 'tej, neidio' -->
  <vowel ph="ei" vc="+" vln="d" vheight="1" vfront="1" vrnd="-" />

  <!-- ##### -->
  <!-- Consonants / Cytseiniaid -->
  <!-- ##### -->
  <!--

  <!-- b /b/ voiced stop/plosive bilabial -->
  <consonant ph="b" vc="-" ctype="s" cplace="l" cvox="+" />

  <!-- k /k/ voiceless velar stop/plosive -->
  <consonant ph="k" vc="-" ctype="s" cplace="v" cvox="-" />

  <!-- x /x/ voiceless uvular fricative -->
  <consonant ph="x" vc="-" ctype="f" cplace="v" cvox="-" />

  <!-- d /d/ voiced alveolar stop/plosive -->
  <consonant ph="d" vc="-" ctype="s" cplace="a" cvox="+" />

</allophones>

```

Datblygu Tagiwr rhannau ymadrodd a Chorpws o Frawddegau Hyfforddi ar gyfer y Gymraeg

GRUFFUDD PRYS

Uned Technolegau Iaith, Prifysgol Bangor, Cymru

GARETH WATKINS

Uned Technolegau Iaith, Prifysgol Bangor, Cymru

Mae'r bennod hon yn cyflwyno'r gwaith cynnar a wnaethpwyd wrth greu TagTeg, tagiwr rhannau ymadrodd Cymraeg newydd sydd wedi'i greu gan ddefnyddio llyfrgell meddalwedd spaCy. Mae adnodd o'r fath yn un o gydrannau pwysicaf a mwyaf sylfaenol unrhyw biblinell prosesu iaith naturiol (NLP). Mae'n cyflwyno'r corpws o frawddegau tagieidig a ddefnyddiwyd i hyfforddi TagTeg hefyd, sydd yn adnodd gwerthfawr a phwysig ynddo'i hun. Mae'r papur yn cyflwyno'r ystyriaethau technegol a ieithyddol roedd yn rhaid ymgymryd â nhw wrth greu TagTeg a'r corpws o frawddegau tagiedig, ac yn trafod y broses o werthuso TagTeg cyn myfyrio ar ganlyniadau'r gwerthuso hynny. Yna, gan mai cyflwyno'r gwaith cynnar yn unig yw diben y bennod hon, cyffyrddwn ar ein cynlluniau ar gyfer y camau nesaf yn natblygiad TagTeg.

Allweddeiriau: tagiwr, corpws, technolegau iaith, Cymraeg, prosesu iaith naturiol (NLP)

1 CYFLWYNIAD

Disgrifia'r bennod hon iteriadau cychwynnol tagiwr rhannau ymadrodd, sy'n un o'r cydrannau prosesu iaith naturiol (NLP) a gynhyrchwyd yn ystod 7 mis cyntaf project Technoleg Testun, Lleferydd a Chyfieithu ar gyfer yr Iaith Gymraeg 2020 – 2021 gan Uned Technolegau iaith, Prifysgol Bangor, dan nawdd Llywodraeth Cymru. Dylid nodi ein bod ni yn nyddiau cynnar y project o hyd, ac er bod ein cynnydd wedi bod yn sylweddol, mae mwy o waith i'w wneud. Diben y bennod hon yw adrodd ar ein cynnydd cynnar a manteisio ar y cyfle i rannu ein profiadau a manteisio ar adborth ein cymheiriaid yn Symposiwm Technolegau Iaith Cymru ym mis Medi 2020.

Yn Adran 2, cyflwynwn y gwaith ar tagiwr rhannau ymadrodd Cymraeg ystadegol a fydd yn ffurfio rhan o biblinell prosesu iaith naturiol yn bennaf ar sail dulliau niwral. Awn ati i ddisgrifio beth yn union yw tagiwr rhannau ymadrodd cyn esbonio'r cyfiawnhad dros greu un newydd sbon ar gyfer y Gymraeg. Trafodwn a chymharwn ddau fath o dagwyr: un math sy'n tagio ar sail rheolau, a math arall sy'n defnyddio dulliau ystadegol. Yn Adran 3 disgrifiwn y broses o greu'r tagiwr, gan gynnwys trafod y dechnoleg yr ydym yn ei defnyddio. Disgrifiwn hefyd y corpws o frawddegau a gasglwyd ac a anodwyd gyda thagiau rhannau ymadrodd er mwyn hyfforddi'r tagiwr hwnnw. Disgrifiwn y broses dagio, y set tagiau, a rhai o ystyriaethau ieithyddol y gwaith. Awn yn ein blaenau wedyn i adrodd ein canlyniadau cychwynnol, cyn disgrifio'r camau nesaf yn natblygiad yr offer. Yn olaf, crynhown ein casgliadau yn Adran 4.

2 TAGTEG: TAGIWR RHANNAU YMADRODD YSTADEGOL CYMRAEG

Yn yr adran hon, disgrifiwn TagTeg, tagiwr rhannau ymadrodd ystadegol a gynhyrchwyd gennym ar sail llyfrgell meddalwedd spaCy.¹

2.1 Beth yw Tagiwr Rhannau Ymadrodd?

Meddalwedd sy'n nodi i ba ddosbarth geiriol y mae gair mewn darn o destun yn perthyn iddo yw tagiwr rhannau ymadrodd. Er enghraifft, byddai disgwyl i dagiwr ddadansoddi testun byr fel 'cath fawr' drwy nodi mai enw (hynny yw, 'noun') yw *cath* ac mai ansoddair yw *fawr*.

CATH/ENW

FAWR/ANSODDAIR

Mae rhai o'r dosbarthiadau geiriol hyn, fel enw, ansoddair a berf, yn gyfarwydd i'r mwyafrif ohonom. Fodd bynnag, gall rhai o'r dosbarthiadau geiriau eraill a ddefnyddir gan rai tagwyr, fel y banodolion (determiners), fod yn llai cyfarwydd, ac nid ydynt o reidrwydd yn cael eu cydnabod gan bob tagiwr neu fframwaith ieithyddol. Gall y ffin rhwng rhai o'r dosbarthiadau geiriol hyn amrywio o dagiwr i dagiwr. Er enghraifft, gallai un tagiwr beidio cydnabod dosbarth geiriol y banodolion a thagio rhai ohonynt fel ansoddeiriau ac eraill fel rhagenwau. Mae'r modd y bydd tagiwr yn tagio geiriau felly yn ddibynnol ar y canllawiau a bennwyd i'r tagiwr eu dilyn, ac ar y set o dagiau mae tagiwr penodol yn ei harddel, heb sôn am gywirdeb technegol y tagiwr ei hun. Golyga hyn fod cymharu cywirdeb gwahanol dagwyr sy'n defnyddio canllawiau a thagsetiau gwahanol yn anodd (am ragor ar hyn gweler [1]).

2.2 Cyfiawnhau Creu Tagiwr Rhannau Ymadrodd Newydd

Yn achos y Gymraeg, mae nifer o dagwyr rhannau ymadrodd eisoes yn bodoli. Datblygwyd un cynnar er mwyn cynorthwyo gyda thagio'r corpws Cymraeg miliwn o eiriau cyntaf, sef CEG [2] Disodlwyd y rheolau tagio a gaed yn y fersiynau cynnar o Cysill gan fodiwl tagio rhannau ymadrodd penodol ar gyfer fersiynau diweddaraf [3]. Yn 2015, rhoddwyd y tagiwr hwn ar gael fel gwasanaeth API rhad ac am ddim [4]. Yn 2010, datblygwyd tagiwr gan Kevin Donnelly, sef Autoglosser er mwyn tagio corpws amlieithog Siarad Prifysgol Bangor [5]. Erbyn 2018 rhyddhawyd fersiwn newydd ohono sy'n canolbwyntio ar yr iaith Gymraeg, sef Autoglosser2 [6]. Defnyddiwyd Eurfa [7], sail geirfaol Autoglosser, gan dimau eraill fel y sail eirfaol ar gyfer eu tagwyr rhannau ymadrodd hwy. Mae'r rhain yn cynnwys WNLT [8] a'i olynydd WNLT2 [9] gan dîm o Brifysgol De Cymru (a ariannwyd gan Lywodraeth Cymru), a CyTag [10] y tagiwr a ddatblygwyd gan dîm o Brifysgol Caerdydd fel rhan o'u project CorCenCC [11], project a ariannwyd gyda grant gan yr AHRC a'r ESRC.

Un nodwedd sy'n gyffredin rhwng y tagwyr hyn i gyd yw mai tagwyr ar sail rheolau ydynt. Bellach, fodd bynnag, tagwyr ystadegol yw'r tagwyr sy'n perfformio orau o edrych ar restr yr ACL o'r tagwyr Saesneg uchaf eu safon yn y maes [12]. Mae'r 8 tagiwr uchaf yn y rhestr yn adrodd cywirdeb o dros 97% ar y set brofi safonol (sef isran benodol o destunau'r Penn Treebank a godwyd yn wreiddiol o'r papur newydd Americanaidd The Wall Street Journal). Mae rhai tagwyr ar sail rheolau hefyd yn adrodd canlyniadau yn y nawdegau uchel. Fodd bynnag, gall y sgoriau uchel hyn fod o ganlyniad i fod wedi teilwra rheolau yn benodol ar gyfer y data gwerthuso. Mae sgôr tagwyr ystadegol yn tueddu i fod yn fwy cynrychioladol o'u gwir berfformiad wrth iddynt

¹www.spacy.io

dagio testunau na welwyd yn ystod y broses ddatblygu, lle ceir strwythurau cystrawennol na chafwyd enghreifftiau ohonynt yn y data gwerthuso.

Yn achos Neale et al. [10], er enghraifft, sy'n nodi i dagiwr CyTag gyrraedd cywirdeb o 96% wrth dagio data a oedd ar gael wrth ddatblygu, mae ein harbrofion ni wedi dangos bod ei berfformiad bywyd bob dydd gyda thestunau nas defnyddiwyd yn ystod y broses ddatblygu yn sylweddol is [1] a bod ffigyrau o gwmpas yr 82% yn fwy cynrychioladol o'i wir berfformiad. Ein cred ni felly wrth fynd ati i greu tagiwr newydd oedd y gallai tagiwr ar sail ystadegau ddarparu canlyniadau a oedd o leiaf 10% yn uwch, ac a fyddai'n sylweddol well am adnabod yr amrywiaeth o enwau pobl, lleoedd, sefydliadau a chynnyrch sy'n codi o fewn testunau cyfoes.

2.3 Gwaith Arall yn y Maes

Ers dechrau ar y gwaith rydym wedi gweld cyfeiriadau at waith ar dagwyr ystadegol Cymraeg gan Heinecke (ar sail UDPIPE)[13], a gwaith gan Ezeani et al.[14] sy'n defnyddio mewnbaniadau geiriau i hyfforddi tagiwr. Nid oes gwerthusiad o system Heinecke ar gael ar hyn o bryd. Mae Ezeani et al. wedi cyhoeddi canlyniadau da, ond nid yw'n ymddangos fel eu bod nhw wedi defnyddio data profi gwahanol i'w data hyfforddi. Byddai cael cyfle i brofi system ar ddata profi annibynnol yn ein galluogi i gymharu eu system nhw gyda TagTeg, fel y gwnaethpwyd yn [1] gyda system CyTag a WNLT2. Fodd bynnag, mae'n ymddangos mai projectau ar waith yw'r rhain ac nad ydynt eto wedi'u pecynnu a'u dosbarthu ar gyfer defnydd cyffredinol ymarferol.

2.4 Tagwyr ar Sail Rheolau a Thagwyr Ystadegol

2.4.1 Tagwyr ar Sail Rheolau

Yn fras, mae tagwyr ar sail rheolau yn pennu rhan ymadrodd gair drwy edrych amdano mewn rhestr gynhwysfawr o eirffurfiau a'u rhannau ymadrodd cyfatebol. Er enghraifft, ar gyfer geirffurf diamwys fel *lleol*, dim ond y rhan ymadrodd *ansoddair* fyddai yn y rhestr ar ei gyfer. Mae ei dagio fel *ansoddair* felly yn ddigon hawdd ac nid oes angen defnyddio rheolau hyd yn oed. Caiff y rheolau eu defnyddio pan fo mwy nag un rhan ymadrodd yn bosib ar gyfer un geirffurf, a bod angen dewis yr un priodol ar gyfer y geirffurf yn y cyd-destun penodol hwnnw.

Er enghraifft, gall y geirffurf *ceir* gynrychioli enw (=noun) lle mae'n cyfeirio at fwy nag un car, neu fe all gynrychioli ffurf amhersonol presennol ar y ferf *cael*. Un ffordd o wahaniaethu rhwng y ddau yw ar sail y geiriau sydd o'u cwmpas. Er enghraifft os mai'r fannod 'r sydd o flaen *ceir*, mae'n debyg mai mwy nag un *car* sydd dan sylw, gan na ddisgwyliad 'r o flaen berf amhersonol. O ysgrifennu llawer o reolau tebyg i hyn, gellir gwella gallu tagiwr i ddewis y tag priodol ar gyfer y geiriau gwahanol sy'n rhannu'r un geirffurf. Mae hyn yn bwysig iawn yn y Gymraeg, gan fod cymaint o eirffurfiau cyffredin yn gallu cynrychioli sawl gair gwahanol. Er enghraifft, gall y gynrychioli'r fannod benodol (*the* yn Saesneg), ond gall hefyd gynrychioli geiryn rhagferfol (fel yn *yr aderyn y gwelsom ei nyth*). Gall y modd y gwiredir treigladau yn orgraffyddol yn y Gymraeg beri i eirffurfiau wrthdaro hefyd, fel yn achos *nos* sy'n gallu cynrychioli ffurf dreigledig ar *dos* fel yn *dyma fy nos cyntaf o'r feddyginiaeth* yn ogystal â'i homograff mwy cyffredin sy'n cyfateb i'r Saesneg *night*. Gyda'r fath gymhlethdod yn yr iaith, problem tagwyr ar sail rheolau yw bod angen i ieithyddion profiadol bentryrru llu o reolau yn ofalus er mwyn ymdrin â holl amrywiadau gramadegol yr iaith tra'n sicrhau nad yw'r rheolau hynny'n ymyrryd â'i gilydd. Heibio rhyw bwynt, mae gwella cywirdeb y tagiwr drwy ychwanegu rheolau yn dod yn fwyfwy annichonadwy.

2.4.2 Tagwyr ar Sail Ystadegol

Mae tagwyr ystadegol yn wahanol gan nad ydynt yn dibynnu ar reolau fel eu prif ddull tagio. Yn hytrach, maen nhw'n gweithredu yn bennaf ar sail y tebygolrwydd bod geirffurf yn perthyn i ran ymadrodd penodol. Yn achos geirffurf fel *ceir*, mae tagiwr ystadegol yn gweithredu trwy bennu'r tebygolrwydd bod *ceir* yn cynrychioli ffurf ar yr enw *car* neu ffurf ar y ferf *cael* yn seiliedig ar ei gyd-destun o fewn y frawddeg lle mae'n digwydd. Mae tagwyr ystadegol yn dysgu model o'r tebygolrwydd hwn drwy gael eu hyfforddi ar frawddegau o destunau lle mae rhan ymadrodd pob gair wedi cael ei anodi â llaw. Er nad yw'r dasg hon o anodi rhannau ymadrodd geiriau â llaw yn dasg syml, tuedda i fod yn haws na cheisio creu, cydbwyso a chynnal cannoedd o reolau gramadegol cymhleth. Mae hefyd yn haws diweddarau'r tagiwr i ymdopi â thestunau problemus gan ei bod yn fater o anodi ac ychwanegu brawddegau enghreifftiol newydd. Yn ymarferol, mae hyn wedi profi'n haws na cheisio pentyrru rheolau gramadegol sy'n gynyddol fwy cymhleth ar ben ei gilydd er mwyn ymdopi gydag eithriadau i reolau gramadegol yr iaith. Un arall o fanteision tagwyr ystadegol yw y gallant ddysgu i gyffredinol yr hyn y maent wedi ei ddysgu o'r broses hyfforddi er mwyn tagio geirffurfiau sy'n ddieithr i'r tagiwr. Gwnânt hynny ar sail nodweddion fel lleoliad gair mewn perthynas â geiriau eraill, ei ragdoddiad a'i olddodiad a phatrymau priflythrennu.

Y prif faen tramgwydd gyda thagwyr rhannau ymadrodd ystadegol yw bod angen anodi llawer o ddata er mwyn hyfforddi tagiwr defnyddiol arno. Cyfeiria Neale et al. [10] at hyn fel un o'r prif resymau dros iddynt ddewis datblygu tagiwr ar sail rheolau yn hytrach na thagiwr ystadegol. Er mwyn creu tagiwr dibynadwy sy'n gallu ymdopi gydag amrywiaeth o fathau gwahanol o Gymraeg (gan gynnwys cyweiriau, tafodieithoedd a phynciau gwahanol) rhaid anodi llawer iawn o destunau amrywiol. I'r tagiwr gael y canlyniadau gorau posib, rhaid i'r testun hyfforddi hyn gynnwys y math o iaith y mae'n fwriad i'r tagiwr ei dagio.

Mae'n rhaid ystyried hawlfraint y data a ddefnyddir hefyd. Hynny yw, os mai'r bwriad yw rhyddhau a dosbarthu'r model, rhaid ystyried a oes gan ddatblygwr yr hawl i ddefnyddio'r data i wneud hynny. Ceir trafodaeth fanwl o hyn gan de Castilho et al. [15] Efallai mai'r peth mwyaf annisgwyl yw eu safbwynt nad yw modelau o reidrwydd yn haeddu cael eu hawlfraint eu hunain gan nad ydyn nhw'n waith creadigol, ac nad ydyn nhw felly yn cyfrif fel gwaith deiliedig hyd yn oed. Fodd bynnag, ymddengys mai'r consensws, gyda modelau iaith, yn ôl spaCy o leiaf, yw bod hawlfraint trwydded y model yn dilyn o'r data [16] Er yr angen i gasglu llawer o ddata a sicrhau ei fod wedi'i drwyddedu yn briodol credwn fod manteision y dull ystadegol o ddefnyddio modelau wedi'u hyfforddi ar ddata yn drech na'r anfanteision.

Mantais ychwanegol ar ben y cywirdeb uwch yw bod modd i sawl tagiwr gwahanol ddefnyddio'r un data hyfforddi. Gan fod y data yn annibynnol i un tagiwr penodol, nid yw'r gwaith o ddatblygu'r data yn mynd yn ofer pan fydd gwell tagwyr ystadegol yn cael eu datblygu (fel sy'n digwydd i reolau tagwyr ar sail rheolau pan fydd gwell tagwyr yn ymddangos). Yn ogystal, mae dull sy'n canolbwyntio ar ddatblygu data tagiedig cyffredinol yn hytrach na rheolau sy'n gaeth i dagiwr penodol hefyd yn osgoi caethiwo datblygu ffyrdd o dagio'r Gymraeg i'r tagiwr arbennig hwnnw. Eisoes y mae nifer o lyfrgelloedd prosesu iaith naturiol y gellid eu hyfforddi ar ddata tagiedig (o bosib gyda rhywfaint o ailfformatio'r data), gan gynnwys llyfrgelloedd John Snow Labs,² spaCy ac UDPipe.³

²<https://www.johnsnowlabs.com/>

³<https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html>

3 CREU'R TAGIWR

Dewiswyd defnyddio tagiwr spaCy fel y sail ar gyfer creu'r tagiwr TagTeg. Llyfrgell brosesu iaith naturiol yw spaCy, sy'n cynnwys piblinell o gydrannau neu 'bibelli' prosesu iaith, gan gynnwys tocyneiddiwr, lemateiddiwr, adnabyddwr endidau enwol a pharsiwr dibyniaethau yn ogystal â'r tagiwr rhannau ymadrodd. Bu i ni ddewis seilio ein gwaith ar spaCy am nifer o resymau. Yn gyntaf, roedd spaCy i weld yn darparu cyfaddawd da rhwng cywirdeb a chyflymder [17] Roedd ambell lyfrgell arall ychydig ganrannau yn fwy cywir, ond tueddai'r cywirdeb ychydig yn uwch hwnnw gostio ar ffurf cynyddu'r amser prosesu yn sylweddol. Ystyriaethau eraill a gyfrannodd at y penderfyniad oedd safon dogfennaeth a chefnogaeth spaCy, a'r ffaith ei fod wedi ei ysgrifennu yn Python (sy'n iaith codio cymharol syml o'i chymharu a ieithoedd eraill ac sy'n hawdd ei darllen a deall) yn yr haenau amlycaf o'r meddalwedd. Yn ogystal, caiff spaCy ei ddosbarthu yn rhad ac am ddim ac o dan drwydded agored ganiataol MIT [18] Rhoddai hyn hyder i ni y byddai'r gwaith yn ddichonadwy ac yn gynaliadwy. Mantais arall i spaCy oedd ei bwyslais ar ddarparu sail ymarferol o safon ddiwydiannol i waith prosesu iaith. Ein cred ni oedd bod hyn yn cyd-fynd yn dda gydag awydd ein harianwyr yn Llywodraeth Cymru i ddarparu offer ymarferol i gefnogi'r defnydd o Gymraeg mewn amgylcheddau digidol fel sgwrsfotiau. Caiff spaCy ei ddefnyddio i ddarparu'r seilwaith prosesu iaith ar gyfer meddalwedd sgwrsfot amlwg Rasa⁴ a meddalwedd anonymieiddio Microsoft Presidio,⁵ er enghraifft. Trwy weithio tuag at greu darpariaeth Gymraeg ar gyfer spaCy ein gobaith oedd y byddai modd i ddatblygwyr meddalwedd cyffredinol ychwanegu'r Gymraeg yn hawdd ac yn hwylus i'r stac meddalwedd y maent eisoes yn bwriadu ei ddefnyddio, a'i fod yn gallu gweithio ochr yn ochr â ieithoedd eraill, gan gynnwys y Saesneg.

3.1 Piblinell spaCy

Fel y nodwyd uchod, cymhelliad arall dros ddatblygu tagiwr newydd Cymraeg oedd yr angen i greu cydran a fyddai'n cymryd ei lle yn dwl fel un 'bibell' brosesu o fewn y biblinell ehangach o gydrannau prosesu iaith a geir o fewn spaCy. Gall y cydrannau hyn fod yn ddibynnol ar ei gilydd. Er enghraifft, o dagio rhan ymadrodd geirffurf fel *ceir* (*cars* neu ffurf ar y ferf *cael*) gellir pasio'r geirffurf a'r rhan ymadrodd i gydran lemateiddio a'i lemateiddio yn briodol i un ai *car* neu *cael* (rhywbeth nad yw WNLT2, er enghraifft, yn ceisio gwneud). O dagio rhannau ymadrodd y geiriau, ceir mwy o wybodaeth y gallai cydrannau pellach y biblinell eu defnyddio, fel yr adnabyddwr endidau enwol sy'n ceisio adnabod endidau fel enwau pobl, sefydliadau a lleoedd, a chydran ar gyfer parsio dibyniaethau, sy'n trosi'r testun o fod yn res o eiriau i fod yn goeden gystrawennol sy'n dangos perthynas y geiriau â'i gilydd – rhywbeth sy'n ddefnyddiol er mwyn deall swyddogaeth gair mewn brawddeg. Er enghraifft, beth yw ystyr 'yn ateb'? Gall gyfateb i yn traethiadol ac enw (e.e. *Mae B yn ateb anghywir*), neu i ategyn berfol a berfenw (e.e. Dafydd oedd **yn ateb** y ffôn). Mae'r cydrannau hyn yn darparu gwybodaeth bwysig ar gyfer prosesau NLP pellach fel systemau parsio bwriad, sgwrsfotiau, systemau crynhoi testun ac ati.

⁴<https://rasa.com/>

⁵<https://github.com/microsoft/presidio>

3.2 Data

3.2.1 Ystyriaethau Data

Rydym eisoes wedi pwysleisio pwysigrwydd data i'r paradeim o hyfforddi modelau iaith ystadegol ar sail data. Er mwyn cael digon o ddata i sicrhau y byddai'r modelau o safon da, nid yw'n fater yn unig o gael mynediad at ddata o faint digonol ag iddo ddigon o amrywiaeth; rhaid hefyd bod â'r hawl cyfreithiol i'w ddosbarthu. Yn ein hachos ni, ein dymuniad oedd gallu dosbarthu nid yn unig y modelau a oedd yn ganlyniad i'r hyfforddi ond y data hyfforddi hefyd, gan y byddai hynny yn galluogi eraill i ychwanegu eu data hwy ato, a hyfforddi modelau mwy teilwredig a mwy eu maint ar gyfer eu dibenion hwy, petaent yn dymuno. Roedd hi hefyd yn fwriad gennym ryddhau'r data o dan drwydded mor ganiataol â phosib. Mae trwyddedau fel CC-BY-SA[19] a'r trwyddedau GPL[20] amrywiol yn caniatáu ailddosbarthu'r data ond yn gosod cyfyngiadau fel yr angen i ddarparu cydnabyddiaeth neu ryddhau unrhyw waith deilliedig o dan yr un drwydded. Ar y llaw arall, mae trwyddedau agored caniatool fel CC0 ac MIT yn rhai hollol agored lle na cheir unrhyw rwystr i'r defnydd o'r data. Mae'r rhain yn fwy derbyniol gan ddiwydiant sydd eisiau defnyddio'r data i greu cynnyrch masnachol yn ddigifyngiad. Gan mai ein nod oedd darparu adnoddau fyddai'n ei gwneud hi'n hawdd a dirwysr i gynnwys y Gymraeg mewn cynnyrch newydd, roedd hon yn ystyriaeth hollbwysig. Penderfynsom felly gasglu data yr oedd modd i ni ei ryddhau o dan drwydded CC0 [21] trwydded sy'n cyfateb i bob pwrpas i osod y data yn y parth cyhoeddus. Lle'r oedd hynny'n bosib, cyfunwyd ein hymdrechion gyda'r gwaith

o gasglu brawddegau i'w cyfrannu at broiect Common Voice Mozilla⁶ ar gyfer eu defnyddio fel promptiau recordio i gasglu recordiadau o wirfoddolwyr yn llefaru brawddegau Cymraeg er mwyn defnyddio'r recordiadau i hyfforddi modelau adnabod lleferydd. Oherwydd hynny, ac oherwydd ein bod yn teimlo y byddai darpar gyfranwyr yn fwy tebygol o gyfrannu brawddegau i'r achos ar ffurf brawddegau unigol cymysg eu trefn yn hytrach na dogfennau cyfan, mae'n well ystyried y casgliad hwn o destunau fel corpws o frawddegau unigol yn hytrach na chorpws o ddogfennau a sgysiau fel corpws CorCenCC. Nid yw'r corpws ychwaith yn un sy'n ceisio casglu ynghyd groestoriad cytbwys o'r iaith. Yn hytrach, y nod yw sicrhau ei fod yn cynnwys digon o enghreifftiau o'r nodweddion y mae model angen eu dysgu wrth gael ei hyfforddi. Yn ogystal â detholiad o'r brawddegau o gorpws Common Voice, mae'r data yn cynnwys brawddegau cyffredinol a awdurwyd gan aelodau'r Uned Technolegau Iaith ym Mhifysgol Bangor, sgysiau ar-lein, a chyfieithiadau o hen lyfrau a storïau Saesneg lle'r oedd yr hawlfraint wedi dod i ben. Ar ben hynny, ceir hefyd erthyglau newyddiadurol, trydariau ac erthyglau gwyddoniadurol a ddarparwyd i ni gan eu hawduron o dan drwydded CC0. Rydym wedi dosbarthu'r ffeil ar ffurf ffeiliau fformat jsonl, lle ceir un cofnod json i bob llinell yn y ffeil, gyda thestun y frawddeg o fewn maes "text", a gwybodaeth amdani (gan gynnwys ei ffynhonnell wreiddiol, manylion trwyddedu a'i id o fewn ein system) yn y maes metadata.

Yn ogystal â 3,345 o frawddegau cyflawn Cymraeg, mae'r set hyfforddi hefyd yn cynnwys 76,097 o 'frawddegau' tagiedig un gair a godwyd o lecsicon yr Uned Technolegau Iaith ac sy'n ffurfio rhan o restr eiriau Cysill a Cysgeir. Defnyddiwyd ein casgliad helaeth o gorpora mewnol i adnabod y geirffurfiau. Cynhwyswyd y brawddegau un-gair hyn er mwyn atgyfnerthu geirfa'r model, ac mae'r model 1% yn fwy cywir o'u cynnwys. Oherwydd natur iaith a'r ffaith fod y geiriau mwyaf cyffredin lawer yn fwy cyffredin na geiriau anghyffredin, yn ymarferol mae hyn yn golygu bod y model yn dangos gwelliant nodedig wrth geisio adnabod geiriau anghyffredin o'i gymharu â model na chafodd

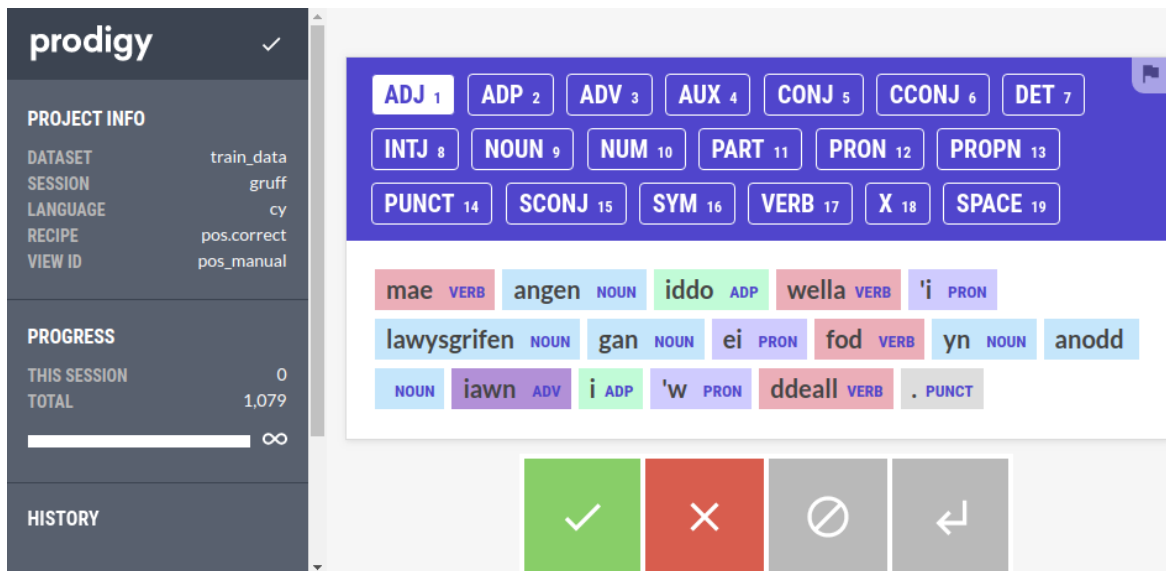
⁶<https://commonvoice.mozilla.org/cy>

ei atgyfnerthu. Gan fod metadata pob brawddeg ynghlwm wrth ei gofnod jsonl, a'r ffynhonnell wedi'i nodi yno, mae modd eithrio'r brawddegau un-gair hyn o'r data pe bai angen.

3.3 Tagio

3.3.1 Offer Tagio

Defnyddiwyd pecyn masnachol Prodigy⁷ gan gwmni explosion,⁸ awduron spaCy, i dagio rhannau ymadrodd y brawddegau a gasglwyd â llaw. Dewiswyd hwnnw gan ei fod yn darparu ffordd hwylus a chyflym o dagio rhannau ymadrodd geiriau'r brawddegau mewn modd a oedd yn gymathus gyda chreu modelau spaCy.



Ffigur 1: Enghraifft o ryngwyneb tagio Prodigy

3.3.2 Set Tagiau

Defnyddiwyd set tagiau Universal Dependencies⁹ wrth dagio gan ei fod yn set drawsieithol syml a chyfarwydd a fyddai'n galluogi cymharu cystrawen y Gymraeg gyda chystrawen ieithoedd eraill, yn arbennig y Saesneg gan fod honno'n iaith a ddefnyddir yn aml ochr yn ochr â'r Gymraeg yng Nghymru.

⁷<https://prodi.gy/>

⁸<https://explosion.ai/>

⁹<https://universaldependencies.org/>

3.3.3 Cysoni'r Tagio

Ceisiwyd harmoneiddio ein dull ni o dagio gyda'r tagiau rhan ymadrodd a roddwyd i eirffurfiau Lecsicon Cymraeg Bangor,¹⁰ sy'n restr gynhwysfawr eirffurfiau Cymraeg a oedd yn cael ei pharatoi i'w rhyddhau o dan drwydded CC0 ar y pryd. Ceisiwyd hefyd sicrhau cysondeb rhwng y ddau dagiwr dynol a fu wrthi. Gwnaethpwyd hynny drwy osod y ddau unigolyn i dagio cyfran helaeth o'r un brawddegau ac adnabod lle y bu iddynt dagio geirffurfiau yn wahanol. Yn sgil hynny, cytunwyd ar ganllawiau ar gyfer tagio'r achosion hynny'n briodol o hynny allan, ac fe aildagiwyd y brawddegau anghyson er mwyn sicrhau cysondeb rhyngddynt. Roedd hi'n ofyniad i ni gyhoeddi'r cydrannau tua 7 mis ar ôl cychwyn y project, felly ni chafwyd amser i sefydlu'r gyfradd cytundeb rhwng anodwyr gwahanol. Y mae'n fwriad gennym i ni wneud hynny yn hwyrach yn y project.

3.3.4 Ystyriaethau ieithyddol

Yn aml, codai'r trafodaethau ynglŷn a sicrhau cysondeb ystyriaethau ieithyddol lle nad oedd datrysiad clir ac unfrydol ar gael yn y llenyddiaeth. Mewn achosion o'r fath, tueddwyd i ffafrio atebion pragmataidd, ymarferol lle y bu'n rhaid i ni ddewis un ffordd dros un arall.

3.3.4.1 Tagio'r berfenw

Un enghraifft o ystyriaeth ieithyddol anodd oedd y penderfyniad (yn wahanol i waith Heinecke [13] ar Universal Dependencies Cymraeg yn eu *Corpws Cystrawenol o'r Gymraeg*) i drin berfenwau fel enwau pan oeddent yn gweithredu'n enwol (fel *canu* yn 'y *canu* da') a'u trin fel berfau pan oeddent yn gweithredu'n fwy berfol (fel *canu* yn 'yn *canu* yn dda'). Gan nad oes tag lefel-uchaf penodol ar gyfer y berfenw o fewn y set drawsieithol a ddefnyddir gan Universal Dependencies, rhaid oedd dewis o blith y tagiau eraill sydd ar gael (gweler <https://universaldependencies.org/u/pos/index.html> am restr o'r rhain). Arfer Heinecke gyda berfenwau yw tagio pob un ohonynt fel enw, ond ein teimlad ni oedd y byddai'r dull a ddewiswyd gennym yn agosach at y consensws academaidd cyfredol, yn cyd-fynd yn agosach â disgwyliadau defnyddwyr cyffredin, ac yn ddefnyddiol er mwyn hwyluso gwahaniaethu rhwng gweithredoedd ar y naill law a phethau a enwir ar y llaw arall. Gobeithiwn gyhoeddi rhagor am hyn yn y man, yn ogystal â chynyddu'r cyd-drafod a chydweithio rhyngom â Heinecke a chyfranwyr ehangach i Universal Dependencies pan fydd amser yn caniatáu.

3.3.4.2 Sgil effeithiau tagio'r berfenw

Un o sgil effeithiau'r penderfyniad i dagio'r berfenw fel berf pan fo'n ymddwyn yn ferfol yw bod hynny'n effeithio ar y dull priodol o dagio'r geiriau hynny sy'n cyfuno ag ef i ffurfio elfennau unigol berfau cwmpasog. O dagio berfenwau fel berfau, penderfynsom hefyd dagio berfau cynorthwyol berfau cwmpasog fel AUX (auxiliary) gan gydnabod y categori o ferfau cynorthwyol fel a wneir mewn dadansoddiadau o ieithoedd eraill o fewn Universal Dependencies, ac o fewn dadansoddiadau ysgolheigion Cymreig fel Thomas [22] o'r Gymraeg. Penderfyniad anodd arall cysylltiedig oedd pennu'r tag priodol ar gyfer geiriau fel 'yn' a 'wedi' a oedd yn wreiddiol fwy na thebyg yn arddodiaid [23] ond sydd bellach yn meddu ar swyddogaeth amser-agwedd-modd (*time-aspect-mood* neu *TAM*). Gan fod y rhain yn amlwg yn cyflawni swyddogaeth wahanol i'r arddodiad, ein hymateb cyntaf oedd eu trin fel PART (*particles* neu eirynnau), ond yn sgil canllawiau Universal Dependencies ar ddynodwyr TAM, efallai y bydd angen ailystyried hynny a'u marcio fel AUX, er bod hynny wedyn yn peri iddynt

¹⁰<https://github.com/techiaith/lecsicon-cymraeg-bangor>

ymddangos fel ei bod nhw'n cyflawni'r union un swyddogaeth â'r berfau cynorthwyol. Mae'n deg dweud hefyd bod i'r penderfyniad rhwng derbyn neu wrthod y posibiliad o dagio berfenw fel berf oblygiadau i gydrannau sy'n codi yn hwyrach yn y biblinell fel y parsiwr dibyniaethau, ac y gallai'r coed cystrawennol a gynhyrchir ar gyfer yr un frawddeg ar sail y ddau ddehongliad gwahanol fod yn wahanol iawn. Mae hon yn ystyriaeth bwysig felly, ac yn haeddu rhagor o ymchwil.

3.3.4.3 Tagio'r DET (determiners)

Enghraifft arall lle nad oedd un cynsail clir ar gael i'w ddilyn oedd wrth geisio gweithredu defnydd Universal Dependencies o'r tag DET (*determiner* neu 'fanodolyn') yn y Gymraeg.

Diffiniad cyffredinol o fanodolyn yw gair sydd yn helpu cyfeirio yn fwy penodol at rywbeth, er enghraifft y *gath*, *rhai cathod*, *dy gath*. Noder fod rhain yn eiriau sydd, y tu allan i gyd-destun ymadrodd neu frawddeg, yn cael eu trin yn draddodiadol fel rhagenwau. Fel dosbarth geiriol, mae banodolyn yn gategori ieithyddol cymharol newydd ym maes ieithyddiaeth. Nid yw'n cael ei gydnabod gan y mwyafrif o eiriaduron Cymraeg, gan gynnwys Geiriadur Prifysgol Cymru. Tuedda hwnnw i labelu'r geiriau fyddai'n gweddu i'r categori hwnnw fel ansoddeiriau dangosol, fel yn achos *hwn*¹¹ Serch hynny, mae gan Thomas[22] ymdriniaeth ohonynt, er y byddem yn dymuno trafodaeth fanylach ohonynt at ein defnydd ni.

Dim ond yn achos y fannod y mae Heinecke yn defnyddio'r tag DET yn ei waith ef ar Universal Dependencies. Teimlwn, fodd bynnag, bod y defnydd ehangach o DET wedi codi am resymau pragmataidd ymarferol, sef yn benodol er mwyn gwahaniaethu rhwng ffurfiau rhagenwau annibynnol fel *rhai* yn *mae hi'n hoffi rhai* a'r *rhai* yn *mae hi'n hoffi rhai pobl* lle mae'r rhai yn digwydd gydag enw i gyfeirio'n fwy penodol neu'n fwy manwl ato. Rydyn ni wedi ceisio dilyn yr arfer ehangach hwnnw gan y teimlwn y byddai gwahaniaethu rhyngddynt yn bwysig er mwyn helpu deall sawl parti gwahanol sydd mewn cymal gan nad yw'r *rhai* yn *rhai pobl* yn cyfeirio at rywrai gwahanol i'r rhywrai y cyfeirir atynt gan y gair *pobl*. Serch hynny, mae'r berthynas rhwng y rhagenwau yn gymhleth ac ystyriwn y bydd angen rhoi sylw pellach i hyn, yn enwedig o ran cadarnhau dull priodol o ddelio â rhagenwau sy'n ailgydio e.e. *dy gath di*.

3.4 Canlyniadau Cychwynnol

3.4.1 Gwerthusiad Meintiol

Gyda'r data a oedd ar gael i ni yn gynnar yn y project rydym eisoes wedi llwyddo i hyfforddi model tagio cychwynnol gyda chywirdeb o dros **91%** ar destun a gasglwyd ar hap ac nad oedd yn y data hyfforddi. Er bod canlyniad cychwynnol o 91% yn ganlyniad cadarnhaol iawn, a'i fod yn ymddangos o'n harbrofion cynnar i fod yn sylweddol yn fwy cywir na nifer o'r tagwyr ar sail rheolau Cymraeg sydd ar gael hyd yma[1] mae hefyd yn awgrymu bod digon o le i wella. Wrth drafod y Saesneg, mae Jurafsky a Martin [24] yn dyfynnu ffigwr o 97% fel yr uchafswm y gellid disgwyl i dagiwr dynol ei gyrraedd, er enghraifft. Os cymerwn bod 500 gair ar un tudalen, byddai cywirdeb o 91% yn dal i olygu bod tua 45 o eiriau wedi'u tagio yn anghywir, ac fe allai'r gwallau hynny fod â goblygiadau sylweddol i dasgau prosesu iaith pellach sy'n ddibynnol ar dagio cywir. Felly, er bod cywirdeb yn y 90au yn ymddangos yn dda, mae angen bod yn ymwybodol o'r safon y mae'r sgoriau hynny'n eu cynrychioli yng nghyd-destun defnydd ymarferol y tagiwr.

¹¹<https://geiriadur.ac.uk/gpc/hwn.html>



Ffigur 2: Enghraifft o allbwn TagTeg

3.4.2 Gwerthusiad Ansoddol

O ran gwerthusiad pellach, unwaith eto, gan ein bod yn adrodd ar y gwaith yn ystod cyfnod cynnar y broses ddatblygu, rhaid bodloni ar hyn o bryd ar ddadansoddiad ansoddol pellach yn hytrach na dadansoddiad meintiol llawn. Bydd gwerthusiad mwy cyffredinol ochr yn ochr â thagwyr eraill yn ymddangos yn [1] a byddwn yn cyhoeddi ystadegau mwy manwl yn hwyrach yn y project. Dyma felly rai o'n prif argraffiadau ni o wendidau'r model tagio cyfredol hyd yma.

Un o ddiffygion amlycaf y tagiwr ar hyn o bryd yw ei fod yn tueddu i dagio pob berfenw fel berf, gan gynnwys y berfenwau y dylai fod wedi'u tagio fel enwau. Mae'n debyg mai'r rhestr o frawddegau un gair sy'n bennaf gyfrifol am hynny, gan y tagiwyd pob berfenw ynddynt fel berfau. Ymddengys nad yw'r nifer o enghreifftiau o ferfenwau enwol o fewn y brawddegau cyflawn yn ddigon i gywiro'r duedd y mae hynny'n ei achosi yn y model. Un ffordd o leihau'r broblem fyddai ychwanegu trigramau tagiedig, er enghraifft *y canu da* a *yn canu yn* yn hytrach na geiriau unigol, felly bwriadwn arbrofi gyda hynny. Mae gennym amheuaeth na fydd y gyfran o ferfenwau enwol a geir mewn brawddegau cyffredin yn ddigon uchel i ddylanwadu'n ddigonol ar y model, felly mae'n bosib y bydd angen atgyfnerthu'r duedd honno o fewn y data hyfforddi, neu o leiaf yn y data a ddefnyddir i werthuso cynnydd cywirdeb y model wrth iddo hyfforddi. Disgwylwn y byddai tagiwr sy'n ceisio gwahaniaethu rhwng berfenw enwol fel *canu* yn '*y canu da*' ac berfenw berfol '*yn canu yn dda*' yn sgorio'n is na thagiwr sy'n trin pob berfenw yr un modd am nad oes yn rhaid iddo ddadamwyso. Serch hynny, awgrymwn na fyddai canlyniadau tagiwr nad ydyw'n dadamwyso yr un mor ddefnyddiol ag un sy'n gwahaniaethu rhwng defnydd enwol a berfol berfenw.

Diffyg arall, ychydig yn annisgwyl, yw bod y tagiwr ambell dro'n adnabod enwau personol fel *Sioned* fel berfau, o bosib dan ddylanwad ffurfiau berfol fel *sonied*. Gobeithiwn y bydd y duedd hon yn diflannu wrth i ni dagio rhagor o frawddegau cyflawn a lleihau'r ddibyniaeth ar y brawddegau un gair.

3.5 Defnyddio'r Tagiwr

Mae fersiwn gychwynnol o TagTeg ar gael i'w lwytho i lawr o <https://github.com/techiaith/model-tag-iwr-spacy-cy>. Gall unrhyw un ddefnyddio'r data i hyfforddi eu model spaCy eu hun, trwy ddefnyddio'r gorchymyn *convert* spaCy ar y llinell orchymyn i drosi'r ffeil i'r fformat hyfforddi cyfredol (gweler <https://spacy.io/api/cli#convert>) ac yna defnyddio gorchymyn *train* spaCy i hyfforddi'r model. I ddefnyddio'r tagiwr, gweler dogfennaeth gyffredinol spaCy.

3.6 Gwaith Pellach

Fel y nodwyd yn adran 1, mae mwy o waith i'w wneud ar TagTeg. Ein bwriad yw casglu mwy o ddata a'i dagio er mwyn nid yn unig cynyddu nifer y brawddegau yn y corpws o frawddegau tagiedig ond hefyd sicrhau gwell amrywiaeth o fewn y data. Rhagwelwn y byddai gwneud hyn, ar ôl i ni ail-hyfforddi'r model ar y corpws chwyddiedig, yn gwella allbwn TagTeg. Fel y nodwyd yn adran 3.4.2, byddwn yn cynnal gwerthusiad llawn o TagTeg, gan gynnwys gwerthusiad trwy ddefnyddio metrigau cywirdeb, adalw ac F1. Yn ogystal â hyn, byddwn yn cynnal gwerthusiad cyfradd cytundeb rhwng anodwyr dynol er mwyn sicrhau cysondeb rhwng yr anodwyr dynol,

ac yna gwirio a chysoni'r anodiadau. Rydyn ni hefyd yn bwriadu cyflwyno ein gwaith i academyddion eraill, gan gynnwys trafod y gwaith gyda'r gymuned ryngwladol sy'n gyfrifol am gychwyn goruchwyllo Universal Dependencies.

4 CASGLIADAU

Yn y bennod hon rydyn ni wedi trafod cynnydd cynnar creu TagTeg, sef tagiwr rhannau ymadrodd ar sail ystadegol ar gyfer y Gymraeg. Ar ôl egluro beth yn union yw tagiwr rhannau ymadrodd, rydyn ni wedi trafod datblygiadau diweddar a hanesyddol yn y maes, ac wedi manylu ar ddau fath o dagiwr gwahanol. Aethom yn ein blaenau i ddisgrifio'r broses o greu TagTeg, gan gynnwys disgrifiad o'r ystyriaethau ieithyddol a hawlfreintiol, a disgrifiad cryno o'r dechnoleg a defnyddiwyd wrth ymafael â'r dasg. Rydyn ni hefyd wedi crynhoi ein canfyddiadau wrth werthuso TagTeg hyd yma.

Mae'n ddyddiau cynnar o hyd, ac mae mwy o waith i'w wneud o ran gwella'r tagiwr, ac o ran darparu fframwaith gwerthuso dyfnach, mwy ystyrlon. Ond o edrych ar ganlyniadau'r gwerthuso hyd yma, gwelwn fod achos gennym i fod yn gadarnhaol o ran gallu TagTeg i dagio testun yn effeithlon ac yn gywir. Dyma gyfraniad pwysig, felly, at y Gymraeg a'i gallu i ymdopi â'r heriau technegol yn y byd sydd ohoni, ac un sydd hefyd yn gosod seilwaith cadarn at y dyfodol.

CYDNABYDDIAETH

Ariannwyd y gwaith hwn gan Lywodraeth Cymru fel rhan o broiect Technoleg Testun, Lleferydd a Chyfieithu ar gyfer yr Iaith Gymraeg 2020 - 2021.

CYFEIRIADAU

- [1] Gruffudd Prys a Gareth Watkins. 2021. Gwerthusiad o Dri Tagiwr Rhannau Ymadrodd Cymraeg. Arxiv.org (ar ddod).
- [2] Nick C. Ellis, C. O'Dochartaigh, William Hicks, M. Morgan a N. Laporte. 2001. Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh. [Ar-lein]
- [3] William Hicks. 2004. Welsh Proofing Tools: Making a Little NLP go a Long Way. Yn Proceeding of the 1st Workshop on International Proofing Tools and Language Technologies. University of Patras, Gwlad Groeg.
- [4] Uned Technolegau Iaith Canolfan Bedwyr Prifysgol Bangor. 2015. API Tagiwr Rhannau Ymadrodd. Cyrchwyd Mawrth 3, 2021 o <https://github.com/PorthTechnolegauIaith/postagger/blob/master/README.md>
- [5] Kevin Donnelly a Margaret Deuchar. 2011. The Bangor Autoglosser: A Multilingual Tagger for Conversational Text. Yn Proceedings of the Fourth International Conference on Internet Technologies and Applications (ITA 11). Prifysgol Glyndwr, Wrecsam, Gogledd Cymru, 17-25.
- [6] Kevin Donnelly. 2018. Autoglosser2 released. Cyrchwyd Medi 1, 2020 o <http://kevindonnelly.org.uk/2018/02/autoglosser2-released/>
- [7] Kevin Donnelly. 2013. Eurfa. Cyrchwyd Chwefror 24, 2021 o <http://eurfa.org.uk>
- [8] Daniel Cunliffe, Doug Tudhope ac Andreas Vlachidis. 2016. The Welsh Natural Language Toolkit WNLT & CYMRIE. Yn WNLT Final Workshop. Prifysgol De Cymru, Trefforest, Cymru.
- [9] Daniel Williams. 2017. Welsh Natural Language Toolkit. Yn WNLT Final Workshop. Prifysgol De Cymru, Trefforest, Cymru.
- [10] Steve Neale, Kevin Donnelly, Gareth Watkins a Dawn Knight. 2018. Leveraging Lexical Resources and Constraint Grammar for Rule-Based Part-of-Speech Tagging in Welsh. Yn Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). LREC, Miyazaki, Japan, 3946-3954.
- [11] Dawn Knight, Steve Morris, Tess Fitzpatrick, Paul Rayson, Irena Spasić ac Enlli-Mon Thomas. 2020. The National Corpus of Contemporary Welsh: Project Report | Y Corpws Cenedlaethol Cymraeg Cyfoes: Adroddiad y Prosiect. Prifysgol Caerdydd<18/>, Cymru.<19/>. arXiv:2010.05542.
- [12] Association for Computational Linguistics. 2020. POS Tagging (State of the art). Cyrchwyd Medi 1, 2020 o [https://adweb.org/adwiki/POS_Tagging_\(State_of_the_art\)](https://adweb.org/adwiki/POS_Tagging_(State_of_the_art))
- [13] Johannes Heinecke a Francis M Tyers. 2019. Development of a Universal Dependencies treebank for Welsh. Yn Proceedings of the Celtic Language Technology Workshop. European Association for Machine Translation, Dulyn, Iwerddon, 21-31.
- [14] Ignatius Ezeani, Scott Piao, Steven Neale, Paul Rayson a Dawn Knight. 2019. Leveraging Pre-Trained Embeddings for Welsh Taggers. Yn Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019): Held at the 57th Annual Meeting of the Association

for Computational Linguistics. Florence, Yr Eidal. Association for Computational Linguistics, Florence, Yr Eidal, 270-280.

- [15] Richard Eckart de Castilho, Giulia Dore, Thomas Margoni, Penny Labropoulou a Iryna Gurevych. 2018. A Legal Perspective on Training Models for Natural Language Processing. Yn Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). LREC, Miyazaki, Japan, 1267-1274.
- [16] Adriane Boyd. 2021. In which license falls a model trained using spacy's LGPL models ? . Cyrchwyd Mawrth 3, 2021 o <https://github.com/explosion/spaCy/issues/7216#issuecomment-787816212>
- [17] Jinho D. Choi, Joel Tetreault, ac Amanda Stent. 2016. It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool. Yn Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Beijing, China, 387-396.
- [18] Fossa. 2021. Open Source Software Licenses 101: The MIT License. Cyrchwyd Mawrth 3, 2021 o <https://fossa.com/blog/open-source-licenses-101-mit-license/>
- [19] Creative Commons. 2021. Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). Cyrchwyd Mawrth 3, 2021 o <https://creativecommons.org/licenses/by-sa/4.0/>
- [20] GNU. 2021. GNU General Public License. Cyrchwyd Mawrth 3, 2021 o <https://www.gnu.org/licenses/gpl-3.0.html>
- [21] Creative Commons. 2021. CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. Cyrchwyd Mawrth 3, 2021 o <https://creativecommons.org/publicdomain/zero/1.0/>
- [22] Peter Wynn Thomas. 1995. Gramadeg y Gymraeg. Gwasg Prifysgol Cymru. Caerdydd.
- [23] Patrick Sims-Williams. 2015. The four types of Welsh YN. Transactions of the Philological Society. 133, 3 (Tachwedd 2015), 286-304. <https://doi.org/10.1111/1467-968X.12052>
- [24] Daniel Jurafsky a James H. Martin. 2020. Speech and Language Processing (3rd ed. draft). Adalwyd Mawrth 3, 2021 o https://web.stanford.edu/~jurafsky/slp3/ed3book_dec302020.pdf

Model Word2vec Cymraeg: cynrychioliad fector o gydberthynas semantig geiriau Cymraeg ar sail eu mewnbaniadau o fewn corpws enfawr Cymraeg

GRUFFUDD PRYS

Uned Technolegau Iaith, Prifysgol Bangor, Cymru

GARETH WATKINS

Uned Technolegau Iaith, Prifysgol Bangor, Cymru

Dyma adrodd ar un o allbynnau cynnar project Technoleg Testun, Lleferydd a Chyfieithu ar gyfer yr Iaith Gymraeg 2020-2021, sef casgliad sylweddol o fectorau geiriau word2vec a hyfforddwyd ar gorpws enfawr o'r Gymraeg ac sy'n ffurfio model o gydberthynas semantig geiriau Cymraeg ar gyfer hwyluso tasgau prosesu iaith. Disgrifiwn y modd y cafodd ei greu a'i ddefnyddioldeb i brosesu iaith, a darparwn werthusiad ansoddol cychwynnol ohono.

Allweddeiriau ac Ymadroddion: Cymraeg, word2vec, corpws, fectorau geiriau, mewnbaniadau geiriau, technolegau iaith, fectorau

1 CYFLWYNIAD

Mae'r bennod hon yn trafod model word2vec Cymraeg sydd o faint sylweddol ac sy'n ddefnyddiol ar gyfer nifer o brosesau NLP gwahanol, yn ogystal â bod yn arf defnyddiol ar gyfer ymchwil geiriadurol a thermolegol. Yn Adran 2, cyflwynwn fewnbaniadau, fectorau geiriau a modelau word2vec gyda chynulleidfa anghyfarwydd mewn golwg. Yn Adran 3 disgrifiwn y data testunol a ddefnyddiwyd i'w hyfforddi ynghyd â'n hymgais ni i gasglu maint sylweddol o destunau Cymraeg a fyddai'n gynhwysfawr ac yn cynrychioli amrywiaeth eang o'r Gymraeg o ran cywair, arddull a maes pwnc. Yn Adran 4 disgrifiwn fanylion technegol y broses hyfforddi. Awn ymlaen yn Adran 5 i ddisgrifio'r canlyniadau, sef y ffurf a gymerodd allbwn y broses hyfforddi. Yn Adran 6 cyflwynwn werthusiad ansoddol o'r model sy'n ceisio amlygu rhai o rinweddau a gwendidau modelau o'r fath ac ystyriaethau perthnasol, yn enwedig rhai sy'n benodol iawn i sefyllfa'r Gymraeg. Yn olaf, yn Adran 7 crynhown ein canfyddiadau ar ffurf casgliadau sy'n amlinellu'r prif bwyntiau a godwyd ac yna nodi'r camau nesaf a ystyriwn yn flaenoriaethau er mwyn datblygu'r gwaith ymhellach.

2 CYFLWYNO MENWBLANIADAU, FECTORAU GEIRIAU A WORD2VEC

Dull o gynrychioli perthynas geirffurfiau â'i gilydd gan ddefnyddio fectorau rhifol yw'r dull word2vec[1] Y nod yw creu cynrychioliad rhifol o eirffurf ar sail ei berthynas â'r geirffurfiau eraill sy'n tueddu i ymddangos o'u cwmpas. Mae'r 'mewnbaniad' hwn yn efelychu gwybodaeth semantig, ac mae wedi profi'n ddull effeithiol o gyfoethogi data testunol gyda gwybodaeth sydd ar y cyfan yn cyd-fynd yn dda â'i ystyron semantig.

Un o brif fanteision dull fel word2vec yw ei fod yn caniatáu trin iaith yn fwy mathemategol. Trwy gynrychioli'r geiriau ar ffurf rhifol, gallwn fesur tebygrwydd gwahanol eirffurfiau, a grwpio geirffurfiau tebyg eu hystyr ynghyd. Gall hyn fod yn ddefnyddiol mewn gwaith geiriadurol, ac ers 2016 mae terminolegwyr yr Uned Technolegau Iaith wedi bod yn defnyddio fectorau word2vec ar sail Corpws Cysill Ar-lein[2] i adnabod geiriau a thermâu tebyg eu hystyr. Mae Palmer a Spasic hefyd wedi gwneud gwaith ar fewnblaniadau yn y Gymraeg fel rhan o broiect Welsh words by numbers: "Wales" + "capital" = "Cardiff" i Lywodraeth Cymru yn 2019 [3]

Un o brif fanteision cynrychioliad rhifol o eirffurfiau yw ei fod yn ein galluogi i drin iaith ar sail mwy cyffredinol. Yn ogystal â bod yn gymorth wrth eiriadura, gall y mewnblaniadau hyn alluogi cyfieithu peirianyddol i ystyried defnyddio geiriau tebyg eu hystyr wrth gyfieithu [4], neu wella gallu tagwyr rhan ymadrodd i ddosbarthu geiriau yn briodol i gategorïau fel enwau, berfau, ansoddeiriau [5] Gyda'r dull word2vec, defnyddir rhwydweithiau niwral i hyfforddi model word2vec ar gasgliadau testun mawr.

Mae'r model a gynhyrchir gan y broses hyfforddi ar ffurf rhestr helaeth o eirffurfiau'r testunhyfforddi (ar wahân i'r rhai mwyaf anghyffredin) ochr yn ochr â'u fectorau cyfatebol. Mae fector pob gair yn cynrychioli perthynas y geirffurf â phob geirffurf arall yn y corpws o destun hyfforddi. Mae'r fectorau hyn yn gweithredu fel cyfeirnodau ar gyfer y geirffurfiau o fewn y gofod fector cysyniadol, fel petaent yn bwyntiau ar fap, gyda'r geirffurfiau sy'n agosaf at ei gilydd o ran y modd y cânt eu defnyddio yn cael eu lleoli yn agosach at ei gilydd o fewn y gofod cysyniadol. Ffurf matrices sydd i'r fectorau hyn o fewn model word2vec. Yn Ffigur 1, dangoswn fector ar gyfer y geirffurf Cymraeg *ci*. Wrth weithredu yn fathemategol ar y fectorau, y matricesau hyn a ddefnyddir.

1.24561942e+00,1.11835730e+00,-8.50843966e-01,-4.85335350e+00,	2.80721927e+01,1.66333973e+00,8.15422118e-01,1.58177555e+00,	-1.67105496e+00,-1.88625574e+00,5.71245611e-01,1.98964465e+00,
-8.02692950e-01,7.72657514e-01,-3.10114026e-01,3.45532149e-01,	-6.70461953e-02,4.26845461e-01,-2.08337620e-01,-8.59721079e-02,	1.01573956e+00,-2.29483986e+00,8.09531808e-02,-9.12274003e-01,
-2.10010959e+00,-4.88950163e-01,-1.70899463e+00,7.55678892e-01,	-7.07006454e-01,-7.97387213e-02,-4.53211099e-01,-8.47332954e-01,	3.72958332e-02,1.03198767e+00,1.25696027e+00,-1.49862409e+00,
-2.15462685e+00,-1.23977207e-01,2.07686639e+00,-3.89734530e+00,	-6.53613448e-01,-2.54306495e-02,-3.09973527e-02,3.58294658e-02,	-1.84663415e-01,-1.13857353e+00,4.83188510e-01,-5.52577615e-01,
-2.26434574e-01,-1.53436506e+00,2.77177151e-01,-1.49922836e+00,	-1.81274557e+00,-1.20999655e-02,-1.19070530e+00,-7.97592044e-01,	-4.20911849e-01,-5.26581109e-01,2.08254278e-01,-1.38960922e+00,
5.90670109e-01,5.65653434e-04,-4.67892319e-01,8.06581080e-01,	-1.41703814e-01,-3.47322583e+00,-2.78286028e+00,-1.81511843e+00,	1.04190586e-02,1.61742222e+00,-3.30863118e-01,-4.71981019e-01,
2.14217782e+00,1.61140430e+00,1.02339971e+00,6.90428764e-02,	5.48577726e-01,5.78229487e-01,-2.37746072e+00,8.71338308e-01,	-1.09296196e+00,-4.87944096e-01,-1.44510770e+00,1.13100362e+00,
-1.40950894e+00,1.66546321e+00,1.53528726e+00,4.09719133e+00,	1.35868752e+00,-3.97622323e+00,2.15092373e+00,-6.96702376e-02,	3.46626163e+00,1.22766078e+00,-4.25548029e+00,6.92818046e-01,
-1.24337220e+00,-2.29404140e+00,-1.75171399e+00,-2.86042869e-01,	2.78959721e-01,1.47650254e+00,1.50279486e+00,5.08396477e-02,	4.73813713e-01,-1.51626170e-02,1.19616259e+00,1.49193823e+00,
-2.87434220e-01,-7.48044729e-01,-3.29672837e+00,3.52492295e-02,	4.21337605e-01,-9.81391549e-01,-6.91332638e-01,-1.37264407e+00,	-4.98459548e-01,1.81485426e+00,-2.80831963e-01,-2.7322219e+00,
-8.95212591e-02,1.10081089e+00,2.77877378e+00,-1.60423446e+00,	-1.10089529e+00,-1.80406404e+00,-2.73046613e-01,1.76466894e+00,	1.67452246e-01,-3.01998162e+00,-9.20212418e-02,2.11629808e-01,
4.31545198e-01,-9.07463908e-01,1.36722767e+00,5.70522308e-01,	3.11397374e-01,4.78620157e-02,4.42940503e-01,7.52618492e-01,	-2.00250649e+00,8.61806154e-01,-1.92957059e-01,-1.41605830e+00,
-2.18986654e+00,-1.17976125e+00,1.28970754e+00,-1.19025660e+00,	1.08891201e+00,-4.92023051e-01,4.67024982e-01,1.66335595e+00,	-1.84123659e+00,-2.22229815e+00,-2.98946595e+00,1.93239439e+00,
5.03311813e-01,1.28839469e+00,-3.18870187e-01,1.72970057e+00,	-2.28662300e+00,2.59271771e-01,1.02383268e+00,-1.51325321e+00,	-1.42395759e+00,1.57196226e+00,-3.17457557e+00,1.83119655e+00,
-1.32465935e+00,1.10540032e+00,-2.45523524e+00,-4.46017236e-02,	-1.92433882e+00,1.35179326e-01,-1.16859585e-01,7.40030944e-01,	-1.23137712e+00,-5.48633993e-01,7.35441625e-01,-9.54666436e-01,
9.50222850e-01,1.70568359e+00,1.32034290e+00,4.38836068e-01,	7.00151503e-01,3.25648808e+00,8.81138265e-01,7.94920683e-01,	8.21753144e-01,5.28462410e-01,-6.92500532e-01,3.50509787e+00,
-1.63671160e+00,-1.49256194e+00,-1.09166527e+00,1.70686162e+00,	-1.85791135e+00,-1.61920011e+00,-1.66247880e+00,2.31383777e+00,	-2.27922177e+00,-8.80043864e-01,-1.44308889e+00,5.42668343e-01,
-1.00236690e+00,-1.50899982e+00,-2.74934709e-01,-2.28433847e+00,	3.40787172e-01,-1.60006726e+00,6.26775503e-01,-3.18726730e+00,	-2.57186353e-01,1.32238722e+00,-5.51940680e-01,-1.49282348e+00,
3.97709489e+00,-1.42622304e+00,-1.64341176e+00,-9.17630792e-01,	6.62129000e-03,-5.48948407e-01,-4.43431854e-01,2.62968659e+00,	-2.46901751e+00,-2.05538440e+00,9.11326945e-01,1.40158677e+00,
9.10295606e-01,3.09330583e+00,1.54324579e+00,-1.96586275e+00,	-1.29215431e+00,-1.62471843e+00,-3.11302233e+00,-6.60264134e-01,	1.61024499e+00,-1.03352222e-02,7.0046568e+00,9.01589811e-01,
7.56567359e-01,8.34176481e-01,-2.55431843e+00,-9.99598503e-02,	-4.24007237e-01,3.05741578e-01,4.14586276e-01,-3.13366020e-01,	-8.44820738e-01,1.01074326e+00,2.48558313e-01,2.94595450e-01,
-3.02684546e-01,3.65683198e-01,1.24553370e+00,4.21393216e-02,	-1.00328147e+00,5.16545283e-01,1.26835060e+00,1.27357912e+00,	-1.45571363e+00,-4.79752511e-01,-2.32537174e+00,6.5346807e-01,
-4.9568480e-01,9.86794412e-01,2.8194213e+00,-2.88773954e-01,	-7.37377107e-01,1.57302594e+00,-2.07202363e+00,1.65039693e+00,	-2.53076673e+00,-4.85277809e-02,1.01159286e+00,-3.44191849e-01,
1.93433434e-01,-8.62822652e-01,3.82833290e+00,8.30399990e-01,	-1.03531468e+00,1.27309954e+00,7.22072244e-01,-1.55807185e+00,	-1.02322853e+00,1.06458807e+00,2.06831169e+00,-2.68436885e+00,
-2.01539898e+00,-2.49486685e+00,-2.03443944e-01,7.49243299e-01,	-2.08869529e+00,-3.66991192e-01,7.18821347e-01,-4.12656248e-01,	2.37540260e-01,-8.53464067e-01,1.35324091e-01,1.35918367e+00

Ffigur 1: Matrics yn dangos fector ar gyfer y geirffurf Cymraeg *ci* o'n model word2vec (noder ei fod wedi ei drefnu ar draws tair colofn uchod er mwyn arbed lle)

Gellir hefyd ddefnyddio norm L2 y matrics i gynrychioli'r fector matrics ar ffurf fwy hylaw y mae modd ei gymharu â'r gwerthoedd ar gyfer geirffurfiau eraill. Er enghraifft, gallai'r norm L2 ar gyfer y geirffurf *athrawes* gael ei gyfleu i ddau le degol fel 0.63, ac mae hynny'n haws i'w drin na'r matrics cyfan. Byddwn yn defnyddio normau L2 wrth werthuso a chymharu tebygrwydd geirffurfiau yn Adran 6.

3 DATA'R CORPWS HYFFORDDI

Os am gynhyrchu model word2vec sy'n darparu fectorau priodol ar gyfer yr iaith Gymraeg yn gyffredinol, rhaid i'r data hyfforddi gynnwys amrywiaeth eang o'r gwahanol fathau o Gymraeg sy'n bodoli ar ffurf testun. Golyga hynny fod angen casglu data sy'n amrywio yn ôl cywair, arddull, tafodiaith, pwnc ac yn y blaen. Mae hynny yn ei dro hefyd yn golygu bod angen casglu llawer iawn o ddata nad yw'n ailadroddus ei natur.

3.1 Casglu Data

Heb os, casglu digon o ddata addas yw prif her hyfforddi model mewnbaniadau geiriau o safon, yn enwedig yng nghyd-destun iaith fel y Gymraeg sydd yn gymharol brin ei hadnoddau. Ers sawl blwyddyn mae'r Uned Technolegau Iaith wedi bod yn ceisio casglu corpora testunol Cymraeg er mwyn hwyluso ymchwil ieithyddol cyfrifiadurol. Cynlluniwyd rhai o'r adnoddau a'r gwasanaethau y buom yn gyfrifol am eu darparu yn benodol fel y gallent ddarparu ffynhonnell o ddata testunol a fyddai'n ein galluogi i wneud defnydd o'r dulliau diweddar sy'n gofyn am lawer iawn o ddata i'w rhoi ar waith. Enghraifft amlwg o hyn yw'r defnydd o wasanaeth Cysill Ar-lein.

3.1.1 Corpws Cysill Ar-lein (334.9 Miliwn Tocyn)

O'r cychwyn cyntaf, datblygwyd Cysill Ar-lein nid yn unig er mwyn darparu gwasanaeth defnyddiol i'r rhai sy'n ysgrifennu yn y Gymraeg, ond hefyd fel ffordd o gasglu corpws sylweddol o destun Cymraeg ar gyfer dibenion ymchwil mewnol yr Uned Technolegau Iaith[2] Gwneir hyn trwy ddarparu gwasanaeth gwirio gramadeg defnyddiol yn gyfnewid am gytundeb y defnyddiwr i gael defnyddio'r data at ddibenion ymchwil. Mae'r testunau hyn yn cynrychioli pob math o Gymraeg fel y caiff ei ysgrifennu heddiw. Serch hynny, credwn fod tuedd gref tuag at destunau o gyd-destun y byd addysg yn y data sy'n adlewyrchu cryfder cymharol y Gymraeg yn yr agwedd honno o fywyd, a'r ffaith fod Cysill Ar-lein yn arf arbennig o ddefnyddiol i'r rhai hynny sy'n astudio ac yn addysgu o fewn byd addysg. Mae'r testunau ar y cyfan yn rai nad ydynt wedi'u golygu ac yn dangos gallu amrywiol o ran safon y Gymraeg ac yn cynnwys mwy o wallau sillafu a gramadeg na thestunau sydd wedi bod trwy broses olygyddol. Yn anffodus, gan fod defnyddwyr gwefan Cysill ar-lein weithiau yn gwirio testunau sydd o natur breifat neu sensitif, nid oes modd i ni rannu'r data hwn mewn fformat lle gellid darllen ei gynnwys. Fodd bynnag, gan fod modelau word2vec yn gynrychioliad cyffredinol bras o berthynas geirffurfiau unigol â'i gilydd, gellir rhannu'r modelau heb rannu gwybodaeth sensitif. Mae hwn yn gorpws monitor sylweddol sy'n cynyddu yn barhaus o ran ei faint wrth ddefnyddio Cysill Ar-lein.

3.1.2 Corpws DECHE (2.1 Miliwn Tocyn)

Enghraifft arall o broiect gyda'r nod o greu corpws yn ogystal ag adnodd defnyddiol mwy cyffredinol oedd y project DECHE[6]. Roedd y project hwn yn atyniadol i'r Coleg Cymraeg Cenedlaethol fel modd o gynnull llyfrgell ddigidol barod o adnoddau ysgolheigaidd Cymraeg ar ffurf e-lyfrau a chyfrolau PDF. Roedd hefyd yn ffordd o gynnull corpws ysgrifennu ysgolheigaidd caboledig Cymraeg sy'n cynnwys gwaith rhai o gewri'r traddodiad ysgolheigaidd Cymraeg, gan gynnwys unigolion fel J. R. Jones a Bobi Jones. Gelir chwilio'r corpws hwn yn www.corpws.cymru.

3.1.3 Corpws Vocab (158.0 Miliwn Tocyn)

Trwy gytundeb gyda sefydliadau sy'n defnyddio gwasanaeth Vocab ar eu gwefannau, casglwyd corpws helaeth o destunau o eiddo'r sefydliadau hynny [7]. Mae'r mwyafrif o'r testunau hyn wedi'u casglu o ffynonellau newyddiadurol cyfoes.

3.1.4 Corpws Addysg (7.2 Miliwn Tocyn)

Corpws o ddeunyddiau addysgol. Casglwyd hwnnw er mwyn hwyluso ymchwil terminolegol yr Uned technolegau Iaith ac yn arbennig y gwaith safoni termau [8].

3.1.5 Corpws Deunyddiau Prifysgol Bangor (33.5 Miliwn Tocyn)

Mae'r corpws hwn yn cynnwys deunyddiau Cymraeg a gynhyrchwyd gan Brifysgol Bangor ac a oedd ar gael i ni, gan gynnwys gwybodaeth o wefan y Brifysgol. Gweithgareddau prifysgol yw'r cynnwys yn fwy na chynnwys pynciol.

3.1.6 Corpws Trydariadau (8.2 Miliwn Tocyn)

Dyma gorpws o drydariadau a gasglwyd gan yr Uned Technolegau Iaith yn 2015[9]. Bu'n rhaid i ni roi'r gorau i ddarparu'r corpws o drydariadau Cymraeg a gasglwyd gennym o Twitter wedi i Twitter newid telerau defnydd eu gwasanaeth. Bellach, mae'n rhaid i'r sawl sydd am ddefnyddio data Twitter lwytho'r data i lawr (yn araf) o'r wefan ei hunain.

3.1.7 Corpws Y Cofnod (21.5 Miliwn Tocyn)

Dyma gorpws sy'n cynnwys segmentau Cymraeg o Gofnod gwahanol 'Gynulladau' cenedlaethol Cymru, gan gynnwys Cynulliad 1999-2003 a Chynulliad 2007-2010 sydd ar gael gan HMSO a'r Cynulliad Cenedlaethol ac sy'n meddu hawlfraint y Goron. Gellir chwilio data'r Cynulladau hynny yn <http://corpws.cymru/ycofnod/>.

3.1.8 Corpws Gwerddon (0.7 Miliwn Tocyn)

Corpws a luniwyd o gyfrolau cylchgrawn academiaidd Addysg Uwch Gwerddon y Coleg Cymraeg Cenedlaethol yw hwn. Mae'n cynnwys trafodaethau academiaidd ar amrywiaeth o bynciau'n ymwneud a'r meysydd Addysg Uwch sy'n cael eu haddysgu ym Mhrifysgolion Cymru. Darpara eirfa dechnegol bwysig o'r meysydd hynny, yn ogystal â chyd-destun technegol pwysig ar eu cyfer.

3.1.9 Corpws Y Gwyddonydd (2.4 Miliwn Tocyn)

Cylchgrawn gwyddonol Cymraeg a redodd o 1963 i 1996 yw Y Gwyddonydd. Casglasom y testunau ohono ar gyfer ymchwil therminolegol yn wreiddiol, gan ei addasu o'r hyn a ddarperir gan Lyfrgell Genedlaethol Cymru fel rhan o'u project Cylchgronau Cymru. Mae'n darparu gwybodaeth bwysig am eirfa wyddonol Gymraeg.

3.1.10 Corpws Beiblau (2.9 Miliwn Tocyn)

Mae'r Beibl yn cynrychioli corpws pwysig ar gyfer llawer o ieithoedd [11]. Mae hynny'n cynnwys y Gymraeg, lle mae sawl fersiwn gwahanol yn bodoli [12]. Defnyddiasom gorpws a luniwyd o gynnwys tri fersiwn gwahanol o'r Beibl, sef yr hyn a adnabyddir bellach fel Beibl William Morgan (sydd mewn gwirionedd yn fersiwn a ddiwygiwyd yn orgraffyddol o'r hyn a ymddangosodd yn 1588), Y Beibl Cymraeg Newydd a ymddangosodd yn 1988, a'r

cyfieithiad cyfoes diweddar, sef beibl.net. Er y gallai cynnwys y Beibl deirgwaith olygu bod rhai geirffurfiau a strwythurau yn cael eu gorgynrychioli, barnasom y dylid eu cynnwys am nifer o resymau gwahanol.

Yn gyntaf, ceir cryn amrywiaeth geirffurfiol rhyngddynt, gyda Beibl William Morgan yn enwedig yn cynnwys ffurfiau berfol cryno sy'n llai cyffredin erbyn hyn ac a fyddai'n anodd eu canfod mewn ffynonellau mwy modern. Penderfynasom ei gynnwys fel y byddai'r model a hyfforddwyd yn gyfarwydd â ffurfiau cryno mwy hanesyddol fel *cychwynasant* a *cychwynnant*, fel yn yr ymadrodd *yn olaf y cychwynnant* (e.e. Numeri 2:31).

O'i gymharu, cynrychiola Y Beibl Cymraeg Newydd arddull sy'n gymharol fwy modern, ond a oedd yn dal i fod yn weddol geidwadol hyd yn oed pan y cafodd ei gyhoeddi yn 1988. Enghraifft o'r duedd hon yw ei fod yn defnyddio *chwi* ar gyfer y rhagenw lluosog ail berson er bod CEG [12] yn dangos bod *chi* yn llawer mwy cyffredin ei ddefnydd tua'r cyfnod hwnnw (1375 *chi* o'i gymharu â 93 *chwi*). Eto, mae arddull y Beibl Cymraeg Newydd yn fwy cwmpasog na Beibl William Morgan. I barhau gyda'r enghraifft o Numeri 2:31, *fydd yr olaf i gychwyn allan* a ddefnyddiwyd ym Meibl 1988 lle cafwyd *yn olaf y cychwynnant* ym Meibl William Morgan.

Erbyn cyhoeddi beibl.net, mae'r gystrawen yn llawer mwy cwmpasog, ac yn fwy rheolaidd felly. Yr hyn a geir yn y fersiwn hwn o'r Beibl yw 'fydd yn symud allan olaf'. Un gwahaniaeth nodedig y sylwasom arno oedd bod nifer y geiriau unigryw yn lleihau yn sylweddol o'r Beibl cynharaf i'r Beibl diweddaraf. Credwn fod hynny oherwydd i'r gystrawen symud o ddefnyddio ffurfiau cryno amrywiol i ddefnyddio berfau cwmpasog lle ceir llai o amrywiaeth yn yr elfennau unigol a ddefnyddir (cymharer *sylwasom* â *rydym wedi sylwi* lle mae *sylw*, *rydym* a *wedi* yn llai tebygol o fod yn unigryw o fewn y testun ehangach na *sylwasom*).

Heb gynnwys y Beibl cynharaf, ein teimlad oedd na fyddem yn llwyddo i gael nifer digonol o fewnblaniadau ar gyfer ffurfiau cryno'r ferf o ffynonellau eraill mwy cyfoes. Teimlwn hefyd na fyddai cystrawen Feiblaidd draddodiadol wedi'i hadlewyrchu yn y mewnblaniadau hynny'n fwy cyffredinol, ac y byddai hynny'n wendid o ystyried dylanwad y Beibl ar yr iaith. Yn sicr, mae anafanteision o fod wedi cynnwys y Beibl deirgwaith yn y data: yr anfantais amlycaf yw bod rhai ffurfiau penodol a geir yn y Beibl wedi'u gorgynrychioli o gymharu â'r gyfradd y disgwyliid eu gweld mewn testun Cymraeg cyffredinol (trafodir hyn yn fanylach yn Adran 6.11). Fodd bynnag, oherwydd yr amrywio o ran cystrawen o un fersiwn i'r llall, ni theimlwn fod hyn yn orbroblemus, yn enwedig gan fod maint y Corpws Beiblau yn ffurfio rhan cymharol fach o'r corpws hyfforddi ehangach.

3.1.11 Corpws Gwyddoniadurol (Wikipedia) (14.3 Miliwn Tocyn)

Crëwyd corpws ar sail allforiad o Wikipedia, y fersiwn Cymraeg o Wikipedia. Teimlwyd ei bod yn bwysig ei gynnwys gan y byddai, fel gwyddoniadur cyffredinol, yn gweithredu fel ffynhonnell werthfawr o wybodaeth semantig gyffredinol. Er hynny, bu'n rhaid treulio llawer o amser yn ei lanhau oherwydd bod cynifer o'r brawddegau ynddo yn amrywiadau ar yr un frawddeg dempled. Er enghraifft, mae'r corpws yn cynnwys cannoedd o frawddegau ar y patrwm *Uchder y copa o lefel y môr ydy [rhif] metr ([rhif] tr)* neu *Cafodd yr uchder ei fesur a'i gadarnhau ar [dyddiad] [mis] [blwyddyn]*. Er bod corpws Wikipedia yn ymddangosiadol fawr cyn y broses ddad-ddyblygu, gwelwyd lleihad sylweddol yn ei faint o ddileu'r brawddegau lled-ddyblyg. Heb y broses ddad-ddyblygu, gallai brawddegau fel yr uchod ddylanwadau yn ormodol ar y model.

3.1.12 Corpws Gwe Enfawr (25.1 Miliwn Tocyn)

Mae'n debyg mai'r 'corpws' mwyaf sylweddol o Gymraeg heddiw yw'r testunau sydd ar gael ar y we. Aethom ati i geisio creu corpws ffurfiol o'r deunyddiau hyn drwy ddefnyddio gwe-ymlusgwr i grafu'r we a chasglu neu 'grafu' testunau, gan geisio targedu testunau Cymraeg yn benodol. Gwnaethpwyd hynny drwy gasglu URLs gwefannau

Cymraeg hysbys i weithredu fel mannau cychwyn crafu, a thrwy eithrio gwefannau lle nad oedd arnynt unrhyw Gymraeg. Ceisiwyd hefyd eithrio gwefannau a oedd yn cynnwys cyfieithu peirianyddol o'r canlyniadau. Cadwai'r crafwr frawddegau unigryw yn unig, sydd â goblygiadau o ran cyfrif mynychder defnydd geiriau, ond oherwydd bod cymaint o frawddegau templed ar y we byddai cadw pob enghraifft o frawddegau cyffredin wedi bod yn fwy problematig. Defnyddiwyd cld2¹ ar yr allbwn er mwyn hidlo brawddegau mewn ieithoedd eraill a oedd wedi eu casglu'n ddamweiniol. Credwn fod y brawddegau sy'n weddill yn y corpws bron i gyd yn rhai Cymraeg dilys. Gobeithiwn barhau i ychwanegu at faint y corpws hwn mewn fersiynau pellach.

3.1.13 Corpora na chynhwyswyd

Am amrywiol resymau, ni chynhwyswyd yn ein corpws hyfforddi ambell gorpws y gellid bod wedi disgwyl i ni fod wedi'u cynnwys. Dewisasom beidio a chynnwys CEG fel y gellid ei ddefnyddio fel corpws profi a chymharu ar gyfer gwerthuso'r corpws hyfforddi yn y dyfodol. Ni chynhwyswyd Corpws Siarad [13] gan y byddai angen bod wedi ailfformio'r data'n sylweddol i gael y testun plaen ohono. Ni chynhwyswyd OSCAR [14] nac unrhyw isrannau Cymraeg o CommonCrawl² gan ein bod yn pryderu y gallai hynny beri dyblygu yn y data. Nid oedd CorCenCC [15] ar gael i ni pan hyfforddwyd y model.

3.2 Glanhau Data

Cyn mynd ati i hyfforddi, roedd angen glanhau'r data er mwyn cael y canlyniadau gorau posib. Roedd hyn yn cynnwys cael gwared â brawddegau dyblyg, yn arbennig rhai sy'n codi o dempledi rhyngwynebau gwefannau, gan gynnwys datganiadau hawlfraint ailadroddus. Mae amgodiad tudalennau gwe yn gallu bod yn broblem fawr, gan nad yw gwefannau wastad yn defnyddio'r amgodiad UTF-8 priodol ar gyfer testunau Cymraeg (sy'n hanfodol ar gyfer cyfleu 'w' ac 'y' yn benodol). Dangosa'r holl broblemau hyn y pwysigrwydd o ddadansoddi unrhyw gorpws a ddefnyddir er mwyn sicrhau nad yw blerwch yn y data yn effeithio yn ormodol ar y canlyniadau.

3.3 Cynnwys y Corpws hyfforddi

Isod ceir trosolwg bras o faint y corpws cyflawn a ddefnyddiwyd i hyfforddi ein model word2vec a'r corpora unigol sy'n gynnwysedig ynddo. Dylid pwysleisio mai ffigyrau bras yw'r rhain gan nad oedd y tocyneiddio terfynol yn orffenedig pan hyfforddwyd y model cychwynnol hwn (bwriadwn ail-greu model newydd unwaith y bydd y tocyneiddio terfynol wedi'i bennu o fewn y biblinell estynedig).

Tabl 1: Cynnwys y Corpora

Corpws	Nifer Tocynnau (y filiwn)
Corpws Cysill Ar-lein	334.9M
Corpws Vocab	158.0M
Corpws Deunyddiau Prifysgol Bangor	33.5M
Corpws Gwe Enfawr	25.1M
Corpws Y Cofnod	21.5M
Corpws Gwyddoniadurol	14.3M

¹<https://github.com/CLD2Owners/cld2>

²<https://commoncrawl.org/>

Corpws	Nifer Tocynnau (y filiwn)
Corpws Cysill Ar-lein	334.9M
Corpws Trydar	8.2M
Corpws Addysg	7.2M
Corpws Beiblau	2.9M
Corpws Y Gwyddonydd	2.4M
Corpws DECHE	2.1M
Corpws Gwerddon	0.7M
Cyfanswm	610.9M

Gyda 610 miliwn o docynnau, dyma'r corpws ymchwil Cymraeg mwyaf y gwyddom ni amdano (er ei bod yn deg dyfalu y byddai mwy na hyn o ddata Cymraeg ym meddiant rhai o'r cwmnïau meddalwedd mwyaf fel Google a Facebook). Credwn ei fod yn cynnwys cynrychioliad da o'r mathau gwahanol o Gymraeg sy'n cael eu cynhyrchu a'u dosbarthu ar y we. Serch hynny, nid yw hynny'n golygu bod iddo gynrychiolaeth gytbwys o'r defnydd cyffredinol o Gymraeg ac nid yw'n cynrychioli ymgais i greu corpws cytbwys o'r Gymraeg fel oedd nod project CorCenCC. Y fantais o beidio ceisio bod yn gytbwys yw ei bod yn haws casglu rhagor o ddata. Adlewyrchir hynny yn y ffaith bod y corpws hyfforddi hwn yn 610M o ran maint tra bod corpws CorCenCC yn 14M. Mae hyn i raddau'n dangos y tyndra sydd rhwng ceisio casglu cymaint o ddata â phosib ar y naill law a cheisio casglu corpws mwy dethol ar y llaw arall.

3.3.1 Ailddosbarthu'r Corpora

Noder nad Prifysgol Bangor sy'n berchen ar hawlfraint y mwyafrif o'r testunau a ddefnyddiwyd i hyfforddi'r model, ac er bod modd i ni ryddhau'r model ar ffurf rhestr o eirffurfiau unigol a'u fectorau, nid oes gennym yr hawl i ailddosbarthu'r mwyafrif o'r corpora hyn gan nad ni yw deiliaid eu hawlfraint. Mewn achosion eraill, fel yn achos Corpws Cysill Ar-lein, mae'r angen i warchod preifatrwydd ein defnyddwyr yn drech na'r ystyriaethau o blaid dosbarthu'r corpws sylweddol hwnnw.

4 HYFFORDDI

Ers cyflwyno word2vec, mae dulliau eraill, mwy soffistigedig o greu mewnbaniadau wedi ymddangos, gan gynnwys GloVe a Fasttext, sy'n ceisio mynd i'r afael â rhai o fannau gwan word2vec. Serch hynny, dechreuasom drwy hyfforddi model word2vec gan mai dyna ar y pryd oedd y fformat mewnbaniadau geiriau a gefnogwyd gan lyfrgell brosesu iaith spaCy, a bod y gwaith hwn yn digwydd law yn llaw gyda'r ymgais i ddatblygu piblinell o bibellau brosesu iaith Cymraeg ar sail spaCy.

Defnyddiwyd pecyn meddalwedd Gensim³ er mwyn hyfforddi'r model, ac roedd y broses honno yn gymharol syml, ac yn llai heriol o lawer na'r dasg o geisio casglu data o faint ac amrywiaeth digonol. Mae Gensim hefyd yn hwyluso hyfforddi modelau Fasttext [16], bwriadwn wneud hynny yn fuan er mwyn gallu cymharu â modelau eraill fel yr un Fasttext parod a ddarperir gan Microsoft.⁴ Disgwylwn i ganlyniad model Fasttext a hyfforddwyd ar yr un data fod yn well na'r model word2vec gan fod hwnnw'n yn galluogi greu mewnbaniadau ar gyfer unedau o dan lefel y gair. Golyga hynny fod modd iddo ddarparu fector ar gyfer geirffurfiau nad ydynt yn yr eirfa (OOV) ar sail cyfuniad o'r elfennau iseiriol (gweler Adran 6.11 am ymdriniaeth lawnach).

³<https://github.com/RaRe-Technologies/gensim>

⁴https://github.com/microsoft/nlp-recipes/blob/master/utils_nlp/models/pretrained_embeddings/fasttext.py

4.1 Gosodiadau Hyfforddi

Wrth hyfforddi'r model gan ddefnyddio Gensim, dewisasom ddefnyddio gwerth o 300 dimensiwn ar gyfer dimensiynau'r fectorau gan y caiff hynny ei ystyried yn ben uchaf yr hyn sy'n arferol [17]., ac y gellid disgwyl iddo felly roi canlyniadau defnyddiol yn ein hachos ni, heb fod y model yn cael ei lyffetheirio gan ddiffyg dimensiynau. Gosodwyd y trothwy ar gyfer y nifer y weithiau y byddai'n rhaid i eirffurf ddigwydd yn y corpws er mwyn cael ei gynnwys yn y model yn 5. Ymddangosai hynny i'n llygaid ni, wedi rhywfaint o arbrofi, yn werth a oedd yn gyfaddawd da rhwng cynnwys geirffurfiau nad oeddent wir yn eirffurfiau dilys ac eithrio geirffurfiau oedd yn rai dilys ond prin.

Gan mai ein bwriad gyda'r gwaith cychwynnol hwn oedd sefydlu gwaelodlin ar gyfer ymchwil arbrofol pellach, cadwyd yn agos at ragosodiadau pecyn Gensim wrth hyfforddi'r model, gan newid y maint fector yn unig o 300 i 100. O ganlyniad, y dull CBOW rhagosodedig a ddefnyddiwyd yn hytrach na'r dull Skip-Gram (gweler <https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314>), ond bwriadwn arbrofi gyda hyn yn y dyfodol.

5 CANLYNIADAU

Canlyniad yr hyfforddi oedd cynhyrchu model a oedd yn cynnwys cyfanswm o 292,769 geirffurf unigryw ynghyd â'u fectorau gair 300 dimensiwn cyfatebol. Penderfynasom chwynnu tocynnau alffaniwmerig fel *abc123* a chymysg eu priflythreniad fel *digwyddiadMawr*, felly nid yw'r o 292,769 geirffurf yn cynnwys tocynnau o'r fath hynny. Ni chafodd y geirffurfiau eu normaleiddio o ran eu priflythreniad, felly mae geirffurfiau fel gwen a Gwen yn gofnodion gwahanol ag iddynt fectorau gwahanol o fewn y data. Mae'r geirffurfiau hyn yn cynnwys geiriau cyffredin Cymraeg yn ogystal ag enwau priod, er enghraifft enwau pobl, lleoedd a chynnyrch etc., yn ogystal â geirffurfiau cyffredin o ieithoedd tramor (o'r Saesneg yn bennaf). Cofnodir y geirffurfiau o fewn y model yn nhrefn eu hamlder yn y corpws, ac mae hynny'n nodwedd ddefnyddiol iawn ar gyfer amrywiaeth o dasgau prosesu iaith.

Tabl 2: Geirffurfiau mwyaf cyffredin a'u Norm L2 cyfatebol

Y tocynnau uchaf eu hamlder yn y corpws hyfforddi (ac eithrio atalnodi a chan gyfuno llythrennau bach a phriflythrennau)	
1. yn	11. wedi
2. i	12. ei
3. y	13. am
4. 'r	14. bod
5. o	15. gan
6. ar	16. mewn
7. mae	17. eu
8. 'n	18. cael
9. yr	19. fod
10. ac	20. hyn

Darparwn y model word2vec a hyfforddwyd ar sail y corpws hyfforddi ar ffurf ffeil *.bin* yn <https://github.com/techiaith/word2vec-cy>, ynghyd â chyfarwyddiadau ar sut i'w ddefnyddio. Mae'r ffeil yn 337Mb o ran maint.

6 GWERTHUSIAD ANSODDOL

Er bod dulliau meintiol o werthuso safon mewnbaniadau wedi cael eu datblygu[18] , nid yw'r cyfle eto wedi codi i baratoi'r data gwerthuso angenrheidiol Cymraeg. Fel y gwerthusiad annibynnol a gomisiynwyd gan Lywodraeth Cymru o'r mewnbaniadau hyn, rhaid i ni ar hyn o bryd droi at werthusiad ansoddol. Dyma eiriau'r gwerthuswr dienw [19]:

The distribution contains a sample script for testing the vectors, including examples on testing several similarity scores for the cases 'athro', 'ysgol' and 'brenin' delivering excellent results.

Rhestrwn isod y canlyniadau hynny er mwyn eu trafod a'u gwerthuso.

6.1 Gwerthusiad 'Athro', 'Disgybl', 'Coleg' ac 'Ysgol'

Isod dangoswn ganlyniad a geir o'r model o ddefnyddio nodwedd *most_similar* Gensim i ganfod y geirffurfiau tebycaf eu hystyr i 'athro' a 'disgybl'. Po uchaf y sgôr tebygrwydd, y tebycaf yw'r geirffurfiau yn ôl y model.

Tabl 3: Geirffurfiau tebyg i *athro*

Geirffurfiau tebyg i athro	Tebygrwydd (Norm L2)
athrawes	0.8889025449752808
ymarferydd	0.737335205078125
addysgwr	0.7125141024589539
athrawon	0.6787909865379333
ymarferwr	0.6760769486427307
tiwtor	0.6219297051429749
hyfforddai	0.6186233758926392
cymhorthydd	0.5997982025146484
Athrawes	0.5851731300354004
ymchwilydd	0.5845403671264648

Tabl 4: Geirffurfiau tebyg i *disgybl*

Geirffurfiau tebyg i disgybl	Tebygrwydd (Norm L2)
dysgwr	0.909702718257904
plentyn	0.7796590328216553
disgyblion	0.7423983812332153
dysgwyr	0.7294760346412659
cyfranogwr	0.710708737373352
myfyriwr	0.6880736351013184
unigolyn	0.6173608303070068
ddisgyblion	0.6116431355476379
plant	0.5878321528434753
cyfranogwyr	0.5865952968597412

Fel y gwelwn yn nhabl 3, mae'r cysyniadau a restrir yn berthnasol iawn i ystyr y gair *athro*. Y cysyniad tebycaf yn y rhestr i *athro* yw *athrawes*, sef athro benywaidd. Mae'n anodd meddwl am gysyniad tebycach gan mai eu rhyw yn unig sy'n gwahanu *athro* ac *athrawes*, ac nad oes i ryw yr addysgwr arwyddocâd arbennig mewn addysg. Mae nifer o'r geirffurfiau tebycaf hyn yn hyponymau neu uwchategori o *athro*. Gwelwn fod *athro* yn fath o *addysgwr* yn ogystal â bod yn fath o *ymarferwr* neu *ymarferydd* addysgol. Mae rhai o'r canlyniadau eraill yn gysyniadau tebyg ond eto ychydig yn wahanol. Gall *tiwtor*, er enghraifft, gynrychioli addysgwr sy'n addysgu un i un, tra bo *cymhorthydd* fel arfer yn addysgwr ychwanegol llai cymwysedig. Ceir patrwm tebyg yn y canlyniadau ar gyfer *disgybl*, sydd hefyd yn berthnasol ac yn ddefnyddiol.

Un nodwedd sy'n dod i'r amlwg yw bodolaeth sawl un o ffurfiau ffurfdroëdig y ffurf sylfaenol, gan gynnwys ffurfiau lluosog ac unigol ar yr un gair. Cawn *athrawon* fel ffurf lluosog ar *athro* neu *athrawes*, er enghraifft. Ceir hefyd y ffurf briflythyrenedig *Athrawes* sy'n cynrychioli'r un cysyniad yn yr achos hwn, ond a allai gyfeirio at gysyniadau gwahanol dro arall, fel yn achos a *glesni* (=glasrwydd) a *Glesni* (enw merch).

Yn y canlyniadau ar gyfer *disgybl*, ceir hefyd yr enghraifft dreigledig *ddisgyblion*. Mae hynny'n amlygu'r ffaith bod i dreigladau cytseiniaid blaen yr ieithoedd Celtaidd oblygiadau wrth greu mewnbaniadau geiriau gan fod y wybodaeth a geir yn y fectorau yn cael eu rhannu a'u gwasgaru rhwng rhagor o eirffurfiau gwahanol na phetai dim

treiglo yn yr iaith. Awgrymwn felly fod y wybodaeth a geir o fewn mewnbaniadau geiriau treiglady yn dlotach ac yn fwy rhanedig (drwy fod wedi'u gwasgaru ar draws nifer o eirffurfiau) na geiriau anhreiglady'r iaith, neu eiriau cyfatebol ieithoedd eraill llai morffolegol gymhleth.

I droi at enghraifft benodol, mae'r wybodaeth ar gyfer holl ffurfiau *disgybl* wedi ei rannu rhwng *disgybl*, *nisgybl* a *ddisgybl* (heb sôn am y ffurfiau lluosog a'r ffurfiau treigledig ohonynt hwythau). Er bod y canlyniadau ar gyfer *disgybl* yn nhabl 4 yn dda, gwelwn nad yw'r rhai ar gyfer *nisgybl* yn nhabl 5 yn ymddangos fel rhai cystal (mae'r rhai ar gyfer *ddisgybl* yn nhabl 6 o bosib yn fwy perthnasol).

Tabl 5: Geirffurfiau tebyg i *nisgybl*

Geirffurfiau tebyg i <i>nisgybl</i>	Tebygrwydd (Norm L2)
nghyfrifiannell	0.8409432768821716
nghanolwr	0.840394139289856
ngywaith	0.8396995067596436
nysgwr	0.8396143913269043
ngweddu	0.838886022567749
nghymhorthydd	0.8387882709503174
nghamddealltwriaeth	0.835710883140564
nghyfnitherod	0.8352251052856445
ngynorthwywyr	0.8346971273422241
ngwahaniaethu	0.8331178426742554

Tabl 6: Geirffurfiau tebyg i *ddisgybl*

Geirffurfiau tebyg i <i>ddisgybl</i>	Tebygrwydd (Norm L2)
fyfyrwr	0.7124826908111572
ddysgwr	0.706892728805542
blentyn	0.6756373643875122
fachgen	0.674723744392395
berson	0.5963745713233948
gystadleuydd	0.5879442691802979
fyfyrwraig	0.5868666172027588
gyfranogwr	0.5710302591323853
gymhorthydd	0.5595928430557251
disgybl	0.5465054512023926

Sylwer hefyd bod y geirffurfiau a geir yn y rhestrau o eiriau tebycaf i air yn tueddu i rannu'r un treigladau â'r gair hwnnw. Yn nhabl 5, mae pob canlyniad yn rhannu'r un treigladau trwynol â'r gair cymharu, sef *nisgybl*. Credwn fod hynny'n dangos cysgod y geiriau rhagflaenol penodol sy'n tueddu i beri treigladau y gair hwnnw. Er enghraifft, bydd y rhagenw 'fy' a'r cysylltair 'neu' yn gyson yn peri i air treigladau sy'n ei ddilyn dreiglo'n feddal [20]. Gan fod mewnbaniadau yn gweithredu ar sail y syniad bod modd diffinio geiriau yn ôl y cwmni y maent yn eu cadw[21] (hynny yw, y geiriau sydd o'u cwmnas), mae'n anorffwrdd y byddai mewnbaniadau geirffurfiau treigledig yn drwm o dan ddylanwad y geiriau cyfagos sy'n ysgogi'r treigladau sydd gan y geirffurf hwnnw. Serch hynny, mae'n syndod bod y patrwm hwnnw yn un mor glir o fewn y canlyniadau ar gyfer *nisgybl*. Teimlwn fod y ffaith mai ffurfiau trwynol fel *nghyfrifiannell* a *ngghanolwr* sy'n dod i'r brig yn awgrymu bod treigladau yn gallu tanseilio ymgais y dull word2vec i weithredu fel modd o glystyru geirffurfiau yn ôl eu hystyron semantig. Yr hyn a welwn wrth geisio adnabod geirffurfiau sy'n debyg eu hystyr i eirffurfiau treigledig yw bod y dull yn hytrach yn amlygu patrymau morffogystrawenol a geir yn nhreigladau'r iaith. Awgryma hyn felly bod angen datblygu dulliau o hyfforddi fectorau sy'n anwybyddu neu o leiaf yn lleihau dylanwad treiglo ar y model.

Un datrysiad posib fyddai dad-dreiglo'r corpws hyfforddi cyn ei ddefnyddio i hyfforddi model, gan fapio'r geirffurfiau treigledig i fewnbaniad y geirffurf di-dreigladau. Hynny yw, gellid troi pob *ddisgybl* a *nisgybl* i *disgybl* yn y data hyfforddi er mwyn cael fector cyfun ar eu cyfer, ac yna gosod y fector cyffredinol hwnnw'n fewnbaniad ar gyfer *disgybl*, *ddisgybl* a *nisgybl*.

Gobeithiwn arbrofi gyda hyn yn y dyfodol gan ddefnyddio tagiwr TagTeg (gweler Prys a Watkins yn y gyfrol hon [22]) sy'n galluogi lemateddio ar sail rhan ymadrodd y gair, ynghyd â'r Offer Trin Iaith a ryddhawyd gennym.⁵ Mae'n bosib y gallai dad-ffurfdroi agweddau eraill fel y ffurfiau lluosog hefyd fod yn ddefnyddiol. Fodd bynnag, gallai addasrwydd dad-ffurfdroi fod yn ddibynnol ar yr union dasg NLP sydd dan sylw. Mae'n bosib y gallai'r treiglo fod yn ddefnyddiol i barsiwr dibyniaethau, er enghraifft. O ganlyniad i hynny bydd yn rhaid arbrofi a gwerthuso hyn. Gallai fod mai'r ateb gorau yw ceisio cynyddu maint y corpws gan obeithio bod canlyniadau ar gyfer *nisgybl* yn gwella drwy hynny, ond mae'n bosib nad oes digon o ddata Cymraeg addas yn bodoli er mwyn gallu gweld gwelliant drwy gynnydd ym maint y data. Gallai gweithrediadau mathemategol i leihau effaith y treigladau hefyd fod yn ddewis arall.

⁵<https://github.com/techiaith/offer-trin-iaith>

Fel gyda'r canlyniadau uchod ar gyfer y ffurfiau didreigladd, mae'r canlyniadau ar gyfer *coleg* ac *ysgol* yn ymddangos yn berthnasol ac yn briodol. Yn achos *coleg* rhestrir nifer o sefydliadau addysgol gwahanol (yn eu plith cawn *feithrinfa*, *ysgol*, *prifysgol*, *llyfrgell* a *clwb*, yn ogystal â'r sesiynau addysgol a gynhelir yno fel *cyrsgau* a *darlithoedd*.

Tabl 7: Geirffurfiau tebyg i *coleg*

Geirffurfiau tebyg i coleg	Tebygrwydd (Norm L2)
brifysgol	0.7886645197868347
goleg	0.6492246985435486
prifysgol	0.6213444471359253
Brifysgol	0.6134623289108276
Coleg	0.6103212833404541
cwrs	0.5707718133926392
campws	0.5558643341064453
colegau	0.5230373740196228
feithrinfa	0.5063283443450928
Goleg	0.5032021999359131
ysgol	0.4937019646167755
cyrsgau	0.49133726954460144
darlithoedd	0.4851536750793457
myfyriwr	0.48434898257255554
llyfrgell	0.4834088385105133
darlithwyr	0.4713546931743622
myfyrwyr	0.4686380624771118
prifysgolion	0.4667356014251709
uwchraddedig	0.46337422728538513
clwb	0.4560936689376831

Tabl 8: Geirffurfiau tebyg i *ysgol*

Geirffurfiau tebyg i ysgol	Tebygrwydd (Norm L2)
Ysgol	0.6604846119880676
ysgolion	0.6279251575469971
ysol	0.5601516962051392
feithrinfa	0.5393844842910767
ysgol-	0.5298383235931396
adran	0.52287358045578
eglwys	0.5205191373825073
athrawes	0.5143859386444092
hysgol	0.504888653755188
coleg	0.4937019646167755
athrawon	0.4809763431549072
athro	0.4706666171550751
ystafell	0.4681328237056732
iard	0.4598560929298401
addysg	0.4504145085811615
ardal	0.4488360285758972
hosbis	0.4486157298088074
amgueddfa	0.44594961404800415
aelwyd	0.4384314715862274
ysbyty	0.4344925284385681

Yn achos *ysgol*, mae'r canlyniadau hefyd yn ymddangos yn briodol o berthnasol i'r cysyniad o sefydliad addysgol neu ganolbwynt cymdeithasol. Mae'r enghraifft o *ysol* yn arbennig o ddiddorol gan ei fod yn amlwg o'r cyd-destun fod y fector yn cynrychioli camsillafiad o *ysgol*. Mae hyn yn amlygu un o broblemau word2vec, sef nad oes modd iddo wahaniaethu rhwng pan fo *ysol* yn gamsillafiad o *ysgol* a phan fo *ysol* yn cynrychioli ansoddair sy'n perthyn i *ysu* ac sy'n golygu *consuming*. Ymddengys yma fod yr ystyr hwnnw wedi cael ei foddi gan berthynas y camsillafiad o *ysgol* â geirffurfiau tebyg o faes addysg.

6.2 Problem Benodol Homograffau Mewn Testun Cymraeg

Nid camsillafiadau fel *ysol* yw'r unig ffynhonnell o wrthdaro rhwng homograffau sy'n rhannu'r un ffurf orgraffyddol ond sy'n cynrychioli cysyniadau gwahanol. Er nad ydynt mor aml yn y Gymraeg ag yn y Saesneg, ceir nifer o eiriau sy'n gallu cynrychioli mwy nag un ystyr. Gelwir y rhain yn homonymau, ac mae *ysgol* yn enghraifft o un ohonynt gan y gall gynrychioli ffrâm o risiau ar gyfer dringo yn ogystal â'r ystyr o sefydliad addysgol. Fel y gwelwn o dabl 8 uchod, mae'r ystyr o ffrâm symudol ar gyfer dringo wedi cael ei boddi'n llwyr gan ganlyniadau addysgol. I raddau, gellid dadlau bod hynny'n adlewyrchu'r ffaith bod yno fwy o ddisgwrs o faes addysg mewn Cymraeg gyfoes

ysgrifenedig na thrafodaeth o offer adeiladu, ond mae hefyd yn tanlinellu bod rhai meysydd yn brinnach eu cynrychiolaeth o fewn y disgwrs Cymraeg cyffredinol.

Yn ogystal â'r amwyster a geir gan ffurfiau cysefin fel *ysgol* sy'n digwydd cynrychiolo mwy nag un cysyniad gwahanol, ceir amwyster pellach yn y Gymraeg oherwydd y modd y gall y treigladau beri i eiriau nad ydynt yn gwrthdaro yn eu ffurfiau cysefin wrthdaro yn ei ffurfiau treigledig [23]. Gall ffurfiau fel *law*, er enghraifft, gynrychioli ffurf dreigledig o *glaw* neu ffurf dreigledig o *llaw*. Yn nhabl 9, gwelwn fod y canlyniadau yn gymysgedd o eirffurfiau sy'n debyg eu hystyr i *glaw* (*gawodydd*, *glaw*, *wynt*, *gymylau*, *lawiad*) a rhai sy'n debyg i *llaw* (*fraich*, *gledr*, *gern* (=boch)).

Tabl 9: Geirffurfiau tebyg i *law*

Geirffurfiau tebyg i <i>law</i>	Tebygrwydd (Norm L2)
gawodydd	0.499844491481781
glaw	0.4749141037464142
gern	0.40383607149124146
fraich	0.39562690258026123
wynt	0.39277803897857666
gledr	0.3897736072540283
forthwyl	0.38594579696655273
gymylau	0.3850449025630951
lawiad	0.38417601585388184
linoleum	0.38245636224746704

Tabl 10: Geirffurfiau tebyg i *is*

Geirffurfiau tebyg i <i>is</i>	Tebygrwydd (Norm L2)
uwch	0.6420556902885437
Is	0.6010016202926636
Iselach	0.5591913461685181
Îs	0.5208320617675781
is-	0.508048415184021
was	0.4835599660873413
uchelach	0.4768733382250366
sub	0.4627269506454468
comes	0.4506301283836365
lower	0.44197165966033936

O fewn cyd-destun y Gymraeg, lle ceir yn aml ymadroddion a theitlau o'r Saesneg, gallai *law* hyd yn oed gynrychioli'r gair Saesneg *law* (=cyfraith. Nid yw'r ystyr honno yn amlwg yn y canlyniadau hyn, ond gwelir dylanwad ystyr geirffurf o'r Saesneg yn glir o chwilio am y geirffurfiau tebycaf i *is* (=lower. Fel y gwelir yn nhabl 10, er mai ffurfiau Cymraeg sydd uchaf yn y canlyniadau, ffurfiau o'r Saesneg yw *was* a *sub* sy'n codi yn chweched ac yn wythfed yn y tabl, gyda *was* yn debyg ei ystyr i'r ferf Saesneg *is*. Er ein bod ni wedi ceisio eithrio testunau Saesneg o'r corpws hyfforddi, mae'n anorod bod rhywfaint o Saesneg yn bwydo i mewn iddo gan fod Saesneg yn codi o fewn brawddegau Cymraeg yn gyffredin ar ffurf teitlau a dyfyniadau.

6.3 Gwerthusiad 'Brenin'

Un o'r enghreifftiau canonaidd [24]. sy'n arddangos defnyddioldeb gallu mewnbaniadau i drosi gwybodaeth eiriol yn wybodaeth rifol lled semantig yw'r un lle dangosir bod modd tynnu gwerth fector y geirffurf *man* o werth fector y geirffurf *king* a chael *queen* fel y geirffurf tebycaf.

Llwyddasom i gael canlyniadau tebyg gyda'n mewnbaniadau Cymraeg drwy dynnu gwerthoedd *dyn* o'r gwerthoedd ar gyfer *brenin* ac ychwanegu'r gwerthoedd ar gyfer *dynes*. Y llinell o god a ddefnyddwyd oedd:

```
show(model.most_similar(positive=['brenin', 'dynes'],negative=["dyn"], topn=10))
```

Ceir enghraifft o hynny yn y ffeil example.py yn ein storfa GitHub.

Yn nhabl 11, gwelwn mai'r tri geirffurf tebycaf i *brenin* yw *Brenin*, *tywysog* a *frenin*. Nid yw *brenhines* yn ymddangos yn y canlyniadau o gwbl. Yn nhabl 12, fodd bynnag, *brenhines* yw'r gair tebycaf, ac mae'r ffurf dreigledig *frenhines* yn ail.

Tabl 11: Geirffurfiau tebyg i *brenin*

Geirffurfiau tebyg i brenin	Tebygrwydd (Norm L2)
Brenin	0.7919192910194397
tywysog	0.7350390553474426
frenin	0.6617879867553711
frenhines	0.6556856632232666
Pharo	0.6539126634597778
brenhinoedd	0.6522496938705444
Herod	0.6422858834266663
pab	0.6362754702568054
Frenin	0.6353232264518738
marchog	0.6326084136962891

Tabl 12: Geirffurfiau tebyg i *brenin* o ychwanegu gwerthoedd *dynes* a thynnu rhai *dyn*

Geirffurfiau tebyg i brenhines – dyn + dynes	Tebygrwydd (Norm L2)
brenhines	0.5559254884719849
frenhines	0.5395671129226685
Pharo	0.5325185656547546
Brenin	0.5258678197860718
brenhinoedd	0.5248633623123169
Dareius	0.4983474612236023
tywysog	0.4887065291404724
Rehoboam	0.4866292178630829
Dál	0.4801225960254669
Antiochus	0.477974534034729

Credwn fod yr enghreifftiau niferus o enwau brenhinoedd ac arweinwyr eraill yn ganlyniad i gynnwys y tri fersiwn gwahanol o'r Beibl yn y data hyfforddi, yn ogystal â Wicipedia, sydd â thuedd i restru enwau brenhinoedd hanesyddol o fewn ei erthyglau. Roedd modd lleihau'r dylanwad hwnnw drwy ddangos y canlyniadau tebyg i *brenin* ond gan dynnu gwerth mewnblianiad *Israel*. Rhoddodd hynny *tywysog*, *frenhines*, *brenhines*, *marchog*, *barwn*, *dywysoges*, *goron*, *porthor*, *telynor*, *frenin* fel y deg canlyniad uchaf (gan gynnwys ffurfiau a briflythrennwyd gyda'r ffurfiau heb eu priflythrennu). Mae'r gallu i ystumio ar y canlyniadau hynny drwy ddulliau mathemategol yn dangos hyblygrwydd y mewnblianiad hyn fel ffordd o gynrychioli geirffurfiau ar ffurf rhifol.

6.4 Gwerthusiad 'Rheoliadau' a 'Firws'

Hyd yma, rydyn ni wedi taro golwg feirniadol ar rai o'r problemau sy'n codi gyda'r dull word2vec yng nghyd-destun y Gymraeg. Rhag tanbrisiu gwerth y model hwn, dyma restru canlyniadau mwy cynrychioladol o'i berfformiad sy'n tanlinellu potensial y dull hwn i dechnolegau iaith Cymraeg. Dewiswyd y geiriau canlynol yn fwy ar hap er mwyn ceisio rhoi blas o'i berfformiad mewn amrywiaeth gwahanol o feysydd.

Er mwyn enghreifftio ei berfformiad gyda thermâu o faes llywodraethiant, dewiswyd y term *rheoliadau* (gweler Tabl 13). Mae'r geirffurfiau a gafwyd yn ôl fel rhai tebyg eu hystyr i gyd yn berthnasol ac yn ymwneud â gosod egwyddorion i'w dilyn o fewn cyd-destun llywodraethu cymdeithas.

Maes arall y mae'r Gymraeg wedi bod yn lledu iddi'n gynyddol yn ystod y blynyddoedd diwethaf yw maes ieuchyd. Gyda'r pandemig diweddar ar flaen ein meddyliau, dewiswyd chwilio'r model am y geiriau tebycaf i *firws*. Roedd y canlyniadau a gafwyd oll yn berthnasol, ac yn amlygu'r ffaith bod dau sillafiad cydnabyddedig ar gyfer cyfateb *virus* yn y Gymraeg, un sy'n mapio yn uniongyrchol o'r ffurf Ladin (sef *firws*), ac un arall (sef *feirws*) sy'n efelychu ynganiad *virus* yn y Saesneg.

Tabl 13: Geirffurfiau tebyg i *rheoliadau*

Geirffurfiau tebyg i rheoliadau	Tebygrwydd (Norm L2)
canllawiau	0.7212351560592651
deddfwriaeth	0.7080656290054321
darpariaethau	0.6954891681671143
deddfau	0.6948240995407104
deddfwriaethau	0.6894554495811462
cyfreithiau	0.6881170272827148
rheolau	0.6874794960021973
rheolaethau	0.6767716407775879
polisiau	0.6742568016052246
reoliadau	0.6652120351791382

Tabl 14: Geirffurfiau tebyg i *firws*

Geirffurfiau tebyg i firws	Tebygrwydd (Norm L2)
feirws	0.7919204235076904
haint	0.741651177406311
firysau	0.6742879748344421
malaria	0.6648149490356445
bacteria	0.6618992686271667
clefyd	0.6611193418502808
bacteriwm	0.656657874584198
pathogen	0.6478495001792908
clwyf	0.6430660486221313
heintiau	0.6409652233123779

Er mwyn ceisio rhoi rhagor o brawf i'r model, aethpwyd ati i geisio dewis geirffurfiau mwy anghyffredin ac astrus.

6.5 Gwerthusiad 'Ebran' a 'Cofalent'

Hen enw ar fwyd neu borthiant ar gyfer ceffylau yw *ebran*. Fe'i dewiswyd er mwyn profi'r model ar eirfa amaethyddol hŷn - y math o beth na ddisgwylir cael llawer o gynrychiolaeth ohono ar y we. Er nad yw *ebran* yn air cyffredin bellach, mae'r canlyniadau fel *cheirch*, *wenith*, *maip* a *rdwins* yn nhabl 15 oll yn berthnasol yng nghydestun bwydydd ar gyfer anifeiliaid, gyda *borant* a *fwytasant* yn cyfleu'r weithred gysylltiedig, a *gascl* (=pentwr) a *tewion* yn eiriau disgrifiadol perthnasol. Er mwyn profi gair o ben arall y sbectrwm, dewiswyd *cofalent*, sy'n disgrifio math o fond cemegol sy'n dal atomau at ei gilydd mewn moleciwl. Mae'r ffurfiau oll yn berthnasol, ar wahân i *onig* sy'n cynrychioli problem amgodio gyda'r ffurf *ionig*. Gellid bod wedi disgwyl gweld *atomau* yn uwch na'i safle (roedd yn 22ain yn y rhestr gyflawn), ond mae'n debyg bod hynny oherwydd mai goleddfwyr ar y cyfan oedd yn cyfateb agosaf i *cofalent* gan mai ansoddair yw *cofalent*.

Tabl 15: Geirffurfiau tebyg i *ebran*

Geirffurfiau tebyg i ebran	Tebygrwydd (Norm L2)
gascl	0.498212069272995
cheirch	0.49560439586639404
tewion	0.49511802196502686
ymenyn	0.49272191524505615
maip	0.49177083373069763
fwytasant	0.47371983528137207
rdwins	0.47243034839630127
wenith	0.4720260500907898
borant	0.47085273265838623
fwsg	0.4696589708328247

Tabl 16: Geirffurfiau tebyg i *cofalent*

Geirffurfiau tebyg i cofalent	Tebygrwydd (Norm L2)
ionig	0.844214677810669
Hydrogen	0.7465869188308716
onig	0.7396742701530457
peptid	0.7186888456344604
gofalent	0.7003843784332275
alcen	0.695235013961792
moleciwl	0.6931548118591309
molecylau	0.6825446486473083
niwclews	0.6779775619506836
hydrogen	0.673660933971405

6.6 Gwerthusiad 'Maswr' a 'Cwrw'

Er mwyn newid cyfeiriad eto, dewiswyd dau derm o feysydd mwy poblogaidd eu naws.

Tabl 17: Geirffurfiau tebyg i *maswr*

Geirffurfiau tebyg i maswr	Tebygrwydd (Norm L2)
mewnwr	0.8384682536125183
blaenasgellwr	0.8043328523635864
cefnwr	0.7888342142105103
bachwr	0.7347344756126404
canolwr	0.7106105089187622
asgellwr	0.6771109104156494
wythwr	0.6747544407844543
blaenwyr	0.6649476289749146
prop	0.6608309745788574
blaenwr	0.6525115966796875

Tabl 18: Geirffurfiau tebyg i *cwrw*

Geirffurfiau tebyg i cwrw	Tebygrwydd (Norm L2)
gwin	0.7391205430030823
seidr	0.7334795594215393
gwirodydd	0.718646764755249
coctels	0.687247633934021
jin	0.6688615083694458
gwrw	0.6675115823745728
chwrw	0.6663733124732971
lager	0.6653351783752441
lemonêd	0.6433229446411133
melysion	0.6420143842697144

Gair o faes rygbi yw *maswr* (=outside half), lle mae'n cynrychioli safle penodol ar y maes rygbi ar gyfer ei chwaraewyr. Roedd y canlyniadau yn berthnasol iawn gan fod pob un ohonynt hefyd yn cynrychioli safle ar y maes.

Y gair nesaf a ddewiswyd oedd *cwrw*, ac unwaith eto roedd y canlyniadau yn berthnasol ac yn perthyn yn bennaf i'r un categori, sef diodydd meddwol, gyda *lemonêd* ac yna *melysion* yn dangos y canlyniadau yn dechrau ymbellhau ohono.

6.7 Gwerthusiad 'Gwarged' a 'Gwaddol'

Dewiswyd hefyd brofi'r canlyniadau ar gyfer *gwarged* a *gwaddol*, dau air sydd fymryn yn anghyffredin, heb iddynt fod yn hollol anghyfarwydd chwaith.

Tabl 19: Geirffurfiau tebyg i *gwarged*

Geirffurfiau tebyg i gwarged	Tebygrwydd (Norm L2)
warged	0.6837567090988159
enillion	0.6060212850570679
refeniw	0.5929793119430542
trosiant	0.5913615822792053
gorwariant	0.5776404142379761
gorbenion	0.5762640237808228
gros	0.5747412443161011
net	0.5730563402175903
elw	0.5707241296768188
drosiant	0.5644094347953796

Tabl 20: Geirffurfiau tebyg i *gwaddol*

Geirffurfiau tebyg i gwaddol	Tebygrwydd (Norm L2)
cyfraniad	0.41669243574142456
etifeddiaeth	0.4150990843772888
cymynrodd	0.41348540782928467
gweledigaethau	0.37813234329223633
waddol	0.3666684329509735
gwreiddiau	0.3642991781234741
creithiau	0.36246055364608765
traddodiadau	0.3608437478542328
Moderniaeth	0.35934978723526
traddodiad	0.35840266942977905

Mae'r canlyniadau ar gyfer *gwarged* (=surplus) oll yn berthnasol i faes cyllid, ond efallai y gellid bod wedi disgwyl gweld *gweddill* yn ymddangos yn uwch yn y canlyniadau. Mae canlyniadau *gwaddol* (=endowment) yn ddiddorol gan eu bod yn adlewyrchu bod y gair yn cael ei ddefnyddio yn drosiadol ar gyfer cysyniadau haniaethol fel

traddodiad a *gwreiddiau* yn ogystal ag ar gyfer etifeddiaeth fwy materol. Mae hefyd yn ddiddorol sylwi bod y canlyniadau ar gyfer *gwaddol* yn briodol iawn er bod y sgôr tebygrwydd yn gymharol isel.

6.8 Gwerthusiad ‘Isio’ ac ‘Eisiau’

Mae mewnbaniadau yn ddefnyddiol iawn yng nghyd-destun testunau sy’n cynnwys llawer o ffurfiau ‘ansafonol’, yn arbennig ffurfiau mwy llafar a thafodieithol. Isod gwelwn ganlyniadau ar gyfer y geiriau tebyg i *isio* a’r geiriau tebycaf i *eisiau*. Noder fod y canlyniadau ar gyfer y ffurf safonol *eisiau* yn cynnwys mwy o ffurfiau safonol tebyg, tra bod y canlyniadau ar gyfer *isio*, sy’n ffurf fwy llafar, yn cynnwys ffurfiau llafar gan fwyaf.

Tabl 21: Geirffurfiau tebyg i *isio*

Geirffurfiau tebyg i <i>isio</i>	Tebygrwydd (Norm L2)
eisio	0.845168948173523
ishe	0.773231029510498
isho	0.7355859875679016
angan	0.6781127452850342
eisiau	0.6601892709732056
licio	0.6232104897499084
eisau	0.6142325401306152
moyn	0.6039586067199707
mhoen	0.5832076072692871
goro	0.5790119171142578

Tabl 22: Geirffurfiau tebyg i *eisiau*

Geirffurfiau tebyg i <i>eisiau</i>	Tebygrwydd (Norm L2)
eisio	0.7607351541519165
isio	0.6601892709732056
angen	0.6397089958190918
dymuno	0.5431622266769409
wedi	0.5423191785812378
ishe	0.5412291884422302
gallu	0.5405140519142151
hoffi	0.5342394113540649
bwriadu	0.5273668766021729
agen	0.5158362984657288

Nid yn unig y mae’r fectorau hyn yn caniatáu i systemau ymdopi gydag iaith lafar, maent hefyd yn adnodd defnyddiol ar gyfer ymchwilio i dafodiaith a normaleiddio ffurfiau llafar.

6.9 Gwerthusiad ‘Anime’, ‘Cartwnydd’ a ‘Cartŵn’

Bu’n her canfod canlyniadau a oedd yn wirioneddol wael ac amherthnasol. Roeddem yn synnu i raddau o weld canlyniadau cystal ar gyfer *anime*, er enghraifft, gan nad oedd hynny’n yn ein taro ni fel pwnc sy’n cael ei drafod yn aml yn y Gymraeg. Serch hynny, un gair lle’r oedd y canlyniadau yn siomedig braid oedd *cartwnydd*, gyda chyfenw’r cartwnydd adnabyddus Huw Aaron a’r term Saesneg cyfatebol *cartoonist* yn uwch na geiriau tebyg eu hystyr.

Tabl 23: Geirffurfiau tebyg i *anime*

Geirffurfiau tebyg i anime	Tebygrwydd (Norm L2)
manga	0.6116302013397217
animeiddiedig	0.5045759081840515
rhaghysbyseb	0.4528508484363556
animeiddio	0.4499656558036804
Avatar	0.4427505433559418
Lolita	0.4420943260192871
Anime	0.4337618350982666
archarwyr	0.43363142013549805
animeiddiadau	0.43353867530822754
Marvel	0.43206268548965454

Tabl 24: Geirffurfiau tebyg i *cartwnydd*

Geirffurfiau tebyg i cartwnydd	Tebygrwydd (Norm L2)
Aaron	0.4788210093975067
cartoonist	0.4635165333747864
Cartŵn	0.4398535192012787
gauntlet	0.43284767866134644
Onllwyn	0.41497212648391724
Bebbiteirawr	0.40730375051498413
ffotograffydd	0.403152734041214
Chiswell	0.3967725932598114
Brassington	0.3945808708667755
gartwnau	0.3921811282634735

Fodd bynnag, roedd y canlyniadau ar gyfer *cartŵn* llawer yn well, ac mae hynny'n dangos ei bod yn anodd darogan lle mae'r diffygion yn y model.

Tabl 25: Geirffurfiau tebyg i *cartŵn*

Geirffurfiau tebyg i cartŵn	Tebygrwydd (Norm L2)
cartwn	0.81761634349823
comig	0.7451732158660889
gartŵn	0.6066957712173462
cartwnau	0.6000196933746338
montage	0.5497390031814575
portread	0.5466684699058533
paentiad	0.5457241535186768
darluniadau	0.5372176170349121
collage	0.5316237211227417

Tabl 26: Geirffurfiau tebyg i *arafu*

Geirffurfiau tebyg i arafu	Tebygrwydd (Norm L2)
cyflymu	0.7991060018539429
gyflymu	0.5817015171051025
stopio	0.5545011758804321
arafi	0.5365293025970459
ymledu	0.5314207077026367
sefydlogi	0.5304709076881409
erydu	0.5254966616630554
gwaethygu	0.5129178762435913
lleihau	0.5124251246452332

6.10 Gwerthusiad 'Arafu'

Un o broblemau cydnabyddedig fectorau geiriau yw'r ffaith bod gwrthystyron fel *hapus* a *trist* yn aml yn cael eu hadnabod fel geiriau tebyg eu hystyr, er eu bod yn cynrychioli dau begwn semantig gwahanol [25]. Yr hyn sy'n achosi'r broblem hon yw bod gwrthystyron yn aml yn digwydd o fewn brawddegau tebyg iawn. Er enghraifft, gellid cyfnewid [gair] mewn brawddeg fel 'rwy'n teimlo'n [gair]' neu 'roedd hi'n [gair] heddiw' am y gair *hapus* neu'r gair *trist* heb i'r frawddeg fod yn un od neu anghyffredin. Yn nhabl 26, uchod, gwelwn enghraifft o'r broblem hyn o fewn y canlyniadau sydd i fod yn debyg i'r geirffurf *arafu* yn y model. Y canlyniad cyntaf (a'r un sy'n bell ar y brig) yw *cyflymu*, sef y gwrthwyneb llwyr i *arafu* o ran ei ystyr. Dyma brawf clir nad gwir ddull semantig yw dull word2vec ond yn hytrach un sy'n dibynnu ar adnabod geiriau sy'n cael eu defnyddio mewn ffordd debyg. Efallai yn ein hachos ni mae'r ffaith fod y corpws yn cynnwys llawer o frawddegau addysgol sy'n trafod arafu a chyflymu mewn modd tebyg a gweddol fformiwläig wedi cyfrannu at amlygrwydd y broblem yma, ond eto mae hon yn broblem gyffredinol hysbys y mae angen bod yn ymwybodol ohoni.

6.11 Geiriau OOV

Un o brif broblemau eraill y dull o ddefnyddio mewnbaniadau yw delio gyda geirffurfiau nad ydynt yn codi yn y model (geiriau 'out of vocabulary' OOV). Fe arbrofasom yn weddol anwyddonol gyda gwahanol eiriau, gan geisio 'baglu' y model drwy gynnig geiriau anghyffredin iddo, ond roedd hi'n anodd meddwl am air Cymraeg dilys nad oedd eisoes wedi'i gynnwys yn y model, ac nid oedd hynny oherwydd diffyg dychymyg ar ein rhan ni. Roedd hyd yn oed geiriau fel *interniaeth*, *hunanyngysu* a *hunlun* nad oeddynt yn ein geiriaduron termau ar y pryd i'w cael o fewn y model. Cawsom fwy o 'lwc' gyda bathiadau newydd sbon o'r gwaith termau, fel *adgyfogi* (=regurgitate) a oedd newydd ei fathu gennym gan fod angen gwahaniaethu rhwng hynny a chyfogi mewn deunyddiau meddygol addysgol.

Er bod y model yn gyffredinol gynhwysfawr felly, mae'n anorod na fydd model word2vec yn cynnwys pob geirffurf posib. Nid yn unig y mae iaith yn newid, ond ceir enwau cynnyrch newydd a chamsillafiadau gwahanol mewn testunau bywyd go iawn lle na fydd mewnbaniadau ar eu cyfer yn y model. Y broblem gyda hyn yw bod fel arfer angen rhyw fath o werth ar gyfer pob gair a geir mewn brawddeg ar gyfer y systemau sy'n gwneud defnydd o fectorau (er enghraifft o fewn systemau cyfieithu peirianyddol niwral). Os nad oes modd cael gwerth priodol o'r model ar gyfer y mewnbaniad, yn aml bydd yn rhaid defnyddio rhyw fath o strategaeth arall i gael gwerth ar gyfer y geirffurf. Er ei bod yn well cael gwerth sydd mor briodol â phosib, mae cael unrhyw werth fel arfer yn well na bod heb werth gan fod angen gwerth i'r system weithio. Mae nifer o strategaethau gwahanol ar gyfer cael mewnbaniad ar gyfer geirffurf lle nad oes mewnbaniad yn bodoli. Gellir defnyddio rhestrau geiriadurol er mwyn ceisio adnabod cyfystyron ac yna defnyddio mewnbaniad y cyfystyr, neu geisio adnabod geirffurfiau sy'n debyg o ran eu sillafiad a defnyddio eu mewnbaniadau hwy. Fodd bynnag, mae terfyn i allu'r dulliau hynny i ddod o hyd i fewnbaniadau lled-briodol. Yr ateb diweddaraf i'r broblem hon yw dulliau sy'n creu mewnbaniadau ar gyfer tocynnau iseiriol (hynny yw, rhannau o eirffurfiau) fel y gellir creu fector ar gyfer geiriau tu allan i'r eirfa (OOV) o gyfuniad ohonynt. Ni fydd hyn o reidrwydd yn creu mewnbaniad hynod o briodol bob tro (nid yw ystyr pob gair yn dryloyw o'i elfennau), ond mae'n gweithio'n dda mewn digon o achosion ac yn bwysicach yn sicrhau bod gwerth ar gyfer pob uned eiriol o fewn y testun. Gobeithiwn roi prawf ar hyn drwy hyfforddi model Fasttext ar sail y corpws hyfforddi yn fuan (mater cymharol syml yw gwneud hynny yn Gensim o fod wedi hyfforddi model word2vec) ac yna cymharu'r canlyniadau.

6.12 Budd i'r Biblinell a Gwaith Termau etc.

Wrth werthuso'r canlyniadau uchod rydyn ni wedi canolbwyntio ar ddadansoddiad o safbwynt ieithyddol dynol. Yn y cyd-destun hwnnw, mae'n glir y byddai'r canlyniadau yn ddefnyddiol ar gyfer ymchwil geiriadurol a therminolegol dynol. Er mwyn ceisio darogan gwerth y model ar gyfer prosesau mwy awtomatig, defnyddiwyd y model i hybu hyfforddi parsïwr dibyniaethau cychwynnol o fewn spaCy. Awgrymai'r canlyniadau y gallai ddarparu 10% o welliant i'r parsïwr hwnnw, ond bydd angen cadarnhau hynny pan fydd y gwaith hwnnw yn fwy datblygedig.

7 CASGLIADAU

I grynhoi, rydym wedi cyflwyno casgliadau cychwynnol sy'n dangos bod modd casglu corpws mawr o destunau Cymraeg a'i ddefnyddio i hyfforddi model word2vec sy'n darparu canlyniadau gwirioneddol ddefnyddiol. Mae ein dadansoddiad ansoddol yn awgrymu ei fod yn gwneud hynny ar gyfer geiriau sy'n perthyn i amrywiaeth o feysydd pwnc a chyweiriau ac arddulliau gwahanol, hyd yn oed pan fo'r ffurfiau yn gymharol anghyffredin. Prin yw'r

achosion lle na cheir fector ar gyfer gair Cymraeg dilys wrth ddefnyddio'r model, ac anaml yw'r canlyniadau amherthnasol.

Fodd bynnag, rydym hefyd wedi adnabod ambell le i wella. Fel y gwelsom, nid yw gwahaniaethu rhwng gwahanol ystyron geirffurfiau yn un o gryfderau modelau mewnblaniadau. Gwelsom hefyd enghraifft o'r broblem gwrthystyr sy'n effeithio ar ddulliau word2vec yn codi ei phen gyda *arafu* a *chyflymu* yn cael eu cyfrif yn debyg. Awgryma hynny fod lle i gynyddu maint y corpws, ac i sicrhau ei fod yn cynnwys brawddegau mwy cyfoethog sy'n gosod y gwrthystyron hyn mewn cyd-destun mwy gwahaniaethol. Gallai hynny olygu canolbwytio ar ychwanegu rhagor o frawddegau llenyddol yn hytrach nag chasglu mwy o destunau gwe a deunyddiau addysgol. Credwn hefyd fod rhagor o le i ychwanegu testunau o feysydd pwnc llai eu cynrychioliad, yn ogystal â chynyddu'r nifer o ysgrifau a gasglwyd sydd mewn arddulliau a chyweiriau mwy prin, heb sôn am gynrychioliad testunol o gyfathrebu llafar Cymraeg. Golyga hyn yn ei dro gasglu rhagor o ddata gan sefydliadau Cymraeg, o gyrff cyhoeddus i weisg preifat, er mwyn gwella'r modelau fydd ar gael i bawb.

Er safon cyffredinol uchel y canlyniadau, rydym wedi adnabod bod treigladau'r Gymraeg yn peri problem ychwanegol i fodelau mewnblaniadau Cymraeg gan ei fod yn gwasgaru gwybodaeth semantig rhwng nifer o ffurfiau ychwanegol lle nad oes gwahaniaeth semantig iddynt mewn gwirionedd. Credwn fod hyn yn golygu bod angen dad-dreiglo neu ddadffurfddroi'r corpws hyfforddi gan ddefnyddio meddalwedd sy'n gallu gwneud hynny mewn modd priodol, gan drosi *ceir* fel berf yn *cael* a *ceir* fel enw lluosog i *car*. Mae ein tagiwr TagTeg yn addawol yn hyn o beth. Bydd angen arbrofi i ganfod pa fathau o ddadffurfddroi sy'n addas, ond ein rhagdybiaeth yw y bydd yn werth chweil ar gyfer tasgau semantig ond efallai'n llai priodol lle bo agweddau cystrawennol i dasg.

Tasg arall sydd i'w gwneud yw arbrofi gydag addasu paramedrau hyfforddi'r model wor2vec, a rhoi cynnig ar hyfforddi model Fasttext er mwyn gweld a geir gwelliant drwy ddefnyddio dulliau mwy soffistigedig o gynhyrchu mewnblaniadau sy'n gallu gweithredu ar sail darnau o eiriau yn hytrach na geirffurfiau cyfan. Serch hynny, ein teimlad cyffredinol yw mai o gynyddu maint ac amrywiaeth y data hyfforddi y cawn y cynnydd mwyaf yn safon y modelau, ac mai gwneud hynny ddylai fod ein blaenoriaeth bennaf.

Gobeithiwn y bydd y gwaith hwn yn ddefnyddiol nid yn unig i hwyluso ymchwil ieithyddol dynol, ond hefyd i wella safon y cydrannau a'r prosesau ieithyddol Cymraeg a allai elwa o'r wybodaeth a geir mewn modelau mewnblaniadau.

CYDNABYDDIAETHAU

Rydym yn ddiolchgar i Lywodraeth Cymru am ariannu'r gwaith hwn fel rhan o broiect Technoleg Testun, Lleferydd a Chyfieithu ar gyfer yr Iaith Gymraeg 2020-2021.

CYFEIRIADAU

- [1] Tomas Mikolov, Kai Chen, Greg Corrad a Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. Yn 1st International Conference on Learning Representations (ICLR 2013). Arizona, UDA, 1-12.
- [2] Delyth Prys, Gruffudd Prys a Dewi Jones. 2016. Cysill Ar-lein: A Corpus of Written Contemporary Welsh Compiled from an On-line Spelling and Grammar Checker. Yn Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Portorož, Slovenia, 3261-3264.
- [3] Geraint I. Palmer. Projects. Cyrchwyd Ebrill 12, 2021 o <http://www.geraintianpalmer.org.uk/projects/>
- [4] Tomas Mikolov, Quoc V. Le, ac Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013).
- [5] Ignatius Ezeani, Scott Piao, Steven Neale, Paul Rayson a Dawn Knight. 2019. Leveraging Pre-Trained Embeddings for Welsh Taggers. Yn Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019): Held at the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Yr Eidal. Association for Computational Linguistics, Florence, Yr Eidal, 270-280.

- [6] Delyth Prys, Mared Roberts a Gruffudd Prys. 2016. Reprinting scholarly works as e-books for less-resourced languages. Yn Proceedings of the Tenth International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), Portorož, Slovenia, 74-79.
- [7] Dewi Jones, Gruffudd Prys a Delyth Prys. 2016. Vocab: A dictionary plugin for websites. Yn Second Celtic Language Technology Workshop. Paris, Ffrainc.
- [8] Tegau Andrews a Gruffudd Prys. 2016. Terminology Standardization in Education and the Construction of Resources: The Welsh Experience. *Education Sciences* 6, 2 (Ionawr 2016), 1-15.
- [9] Dewi Jones, Patrick Robertson, ac A. Taborda. 2015. Corpus of Welsh Language Tweets. *Cyrchwyd Ebrill* 12, 2021 o <http://techiaith.cymru/data/corpora/twitter/?lang=en>
- [10] Thomas Mayer a Michael Cysouw. 2014. Creating a Massively Parallel Bible Corpus. Yn Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). European Language Resources Association (ELRA), Reykjavik, Iceland, 3158-3163.
- [11] Cymdeithas Y Beibl. 2015. Y Beibl Cymraeg. *Cyrchwyd Ebrill* 12, 2021 o <https://www.biblesociety.org.uk/cymru/y-beibl-cymraeg/?cymraeg>
- [12] Nick C. Ellis, C. O'Dochartaigh, William Hicks, M. Morgan a N. Laporte. 2001. Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh. [Ar-lein].
- [13] Kevin Donnelly a Margaret Deuchar. 2011. The Bangor Autoglosser: A Multilingual Tagger for Conversational Text. Yn Proceedings of the Fourth International Conference on Internet Technologies and Applications (ITA 11). Prifysgol Glyndwr, Wrecsam, Gogledd Cymru, 17-25.
- [14] Pedro Javier Ortiz Suárez, Laurent Romary a Benoît Sagot. 2020. A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. Yn Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Ar-lein, 1703-1714.
- [15] Dawn Knight, Steve Morris, Tess Fitzpatrick, Paul Rayson, Irena Spasić, Enlli-Mon Thomas, Alex Lovell, Jonathan Morris, Jeremy Evas, Mark Stonelake, Laura Arman, Josh Davies, Ignatius Ezeani, Steve Neale, Jennifer Needs, Scott Piao, Mair Rees, Gareth Watkins, Lowri Williams, Vignesh Muralidaran, Bethan Tovey-Walsh, Laurence Anthony, Thomas M. Cobb, Margaret Deuchar, Kevin Donnelly, Michael McCarthy a Kevin Scannell. 2020. CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh Dataset. <http://doi.org/10.17035/d.2020.0119878310>.
- [16] Piotr Bojanowski, Edouard Grave, Armand Joulin a Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5, (2017), 135-146.
- [17] Kevin Patel a Pushpak Bhattacharyya. 2016. Towards Lower Bounds on Number of Dimensions for Word Embeddings. Yn Proceedings of the 8th International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing, Taipei, Taiwan, 31-36.
- [18] Bakarov, Amir. 2018. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536* (2018).
- [19] Dienw. 2020. Assessment of Deliverables (Welsh Language Technology). Adroddiad Gwerthuso.
- [20] Martin J. Ball a Nicole Müller. 2002. *Mutation in Welsh*. Routledge.
- [21] John R. Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*: 1-32. Reprinted in F.R. Palmer, ed. 1968. *Selected Papers of J.R. Firth 1952-1959*. Longman, Llundain.
- [22] Gruffudd Prys a Gareth Watkins. 2021. Datblygu Tagiwr Rhan Ymadrodd a Chorpws o Frawddegau Hyfforddi ar gyfer y Gymraeg. Iait h a Thechnoleg yng Nghymru Cyfrol 1. Uned Technolegau Iait h Canolfan Bedwyr Prifysgol Bangor, Bangor, Cymru.
- [23] Martin J. Ball. 1990. Homonymic clash and initial lenition in Welsh: 1,(000),000,000 problems. *Word* 41, 3 (1990), 329-335.
- [24] Levy, Omer, ac Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. Yn Proceedings of the eighteenth conference on computational natural language learning, 2014. Association for Computational Linguistics, Ann Arbor, Michigan, 171-180.
- [25] Bruna Thalenberg. 2016. Distinguishing antonyms from synonyms in vector space models of semantics.

Gwreiddio NMT mewn cwmni cyfieithu Cymraeg

Heriau a chanfyddiadau project i wreiddio technoleg cyfieithu niwral mewn cwmni cyfieithu Cymraeg

MYFYR PRYS,

Cymen Cyf, Caernarfon, Cymru

Mae'r bennod hon yn disgrifio canlyniadau partneriaeth blwyddyn o hyd rhwng Prifysgol Bangor, Llywodraeth Cymru a'r cwmni cyfieithu Cymen Cyf i ddatblygu a gwreiddio technoleg cyfieithu peiriannau niwral yn llifoedd gwaith a seilwaith technolegol y cwmni. Defnyddiwyd y pecyn meddalwedd cod agored Marian NMT i hyfforddi peiriannau cyfieithu niwral y gellid eu cynnwys ym meddalwedd cyfieithu'r cwmni. Mesurwyd ansawdd cymharol y peiriannau cyfieithu hyn gan ddefnyddio'r metrig BLEU, a gwelsom gynnydd o 5 pwynt yn ansawdd y peiriannau ar ôl tiwnio hyperbaramedrau'r model. Mae'r fersiwn wedi'i optimeiddio o'n peiriant yn perfformio 9 pwynt BLEU yn well na Google Translate ar ddata mewnol Cymen ei hun, a 5 pwynt BLEU yn well ar set ddata allanol 'niwtral'. Yn ogystal â datblygu'r peiriannau eu hunain, rhan arall o orchwyl y project oedd datblygu ap i alluogi defnyddio'r dechnoleg mewn offeryn cyfieithu gyda chymorth cyfrifiadur.

Geiriau ac ymadroddion allweddol: cyfieithu peirianyddol niwral, economi Cymru, cofion cyfieithu

1 CYFLWYNIAD

Mae Llywodraeth Cymru wedi nodi'n ddiweddar bod y sector cyfieithu yng Nghymru yn faes hollbwysig i'w ddatblygu [1]. Yn ystod y degawdau diwethaf, gwelwyd twf graddol yn y galw am gyfieithu o'r Saesneg i'r Gymraeg, o ganlyniad i ddwy don o ddeddfwriaeth sydd â'r bwriad o ddiogelu'r Gymraeg [2]. Yn ei phapur strategaeth diweddaraf, Cymraeg 2050: Miliwn o Siaradwyr Cymraeg [1], mae Llywodraeth Cymru yn disgrifio'r angen am "broffesiwn cyfieithu fodern sy'n manteisio i'r eithaf ar dechnoleg, ac adnoddau ieithyddol (yn eiriaduron, terminolegau, a chorpora) [...]". Gellir deall y 'dechnoleg ddiweddaraf' yma yn bennaf fel cyfeiriad at dechnoleg cof cyfieithu a chyfieithu peirianyddol, dau ddatblygiad arloesol sydd â'r potensial i godi cynhyrchedd y cyfieithydd dynol yn sylweddol. Mae ymchwil gan Screen [3] wedi tanlinellu'r potensial o fabwysiadu technoleg newydd i gynyddu cynhyrchiant yn niwydiant cyfieithu Cymru. Mae ei waith ar gofion cyfieithu a chyfieithu peirianyddol yn amcangyfrif y gallai defnyddio technoleg cyfieithu peirianyddol godi cynhyrchedd grŵp o wyth cyfieithydd o 2.5 miliwn o eiriau'r flwyddyn i tua 7 miliwn o eiriau'r flwyddyn.

Mewn papur blaenorol, mae Prys a Jones [4] wedi disgrifio gwreiddio technoleg cyfieithu peirianyddol mewn cwmni cyfieithu yng Ngogledd Cymru, sef Cymen Cyf. Roedd y project hwnnw'n rhan o Bartneriaeth Trosglwyddo Gwybodaeth (KTP) sy'n gyfrwng a ddefnyddiwyd gan Lywodraeth y DU i feithrin cysylltiadau rhwng y sector ymchwil a'r sector breifat. Y rhesymeg dros y bartneriaeth oedd cynyddu cynhyrchiant y cwmni drwy wreiddio peiriannau cyfieithu parth benodol yn ei lifoedd gwaith. Nod arall oedd trosglwyddo gwybodaeth academiaidd i'r cwmni a gwella sgiliau'r staff yn barhaol, gan arwain at effaith gadarnhaol ar y sector yn ei chyfanrwydd. Ystyriwyd bod y project hwnnw'n llwyddiant a dyfarnwyd y radd uchaf bosib iddo gan y corff cyllido, Innovate UK. Mae'r

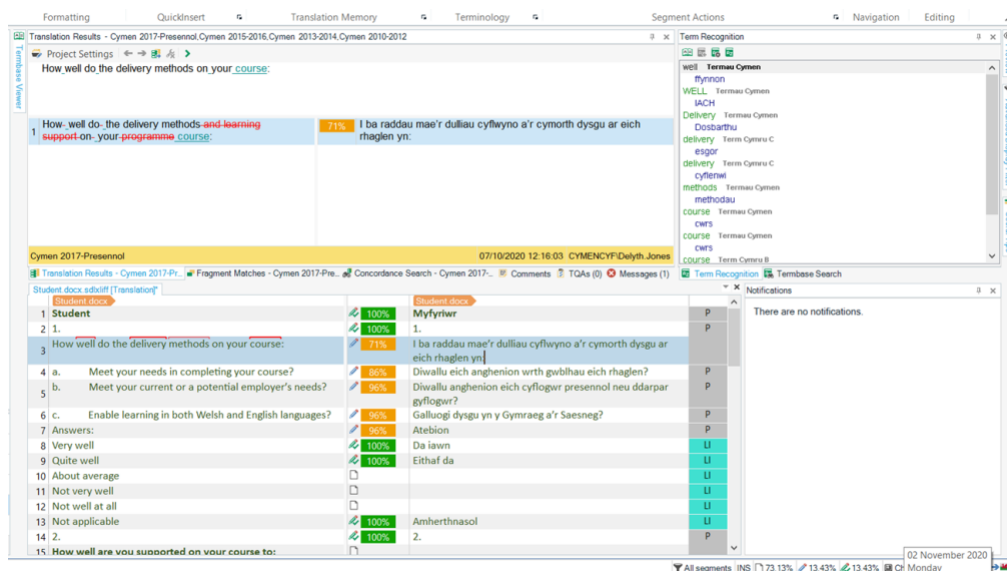
bennod hon yn adrodd ar broiect dilynol, Partneriaeth SMART rhwng Prifysgol Bangor, Cymen a Llywodraeth Cymru. Er bod y Bartneriaeth Trosglwyddo Gwybodaeth wedi rhoi technoleg cyfieithu peirianyddol ystadegol ar waith, mae'r Bartneriaeth SMART newydd wedi canolbwyntio ar uwchraddio i'r paradeim cyfieithu peirianyddol niwral newydd (NMT). Bydd yr adran nesaf yn cyflwyno cyfieithu peirianyddol a thechnoleg cof cyfieithu yn gryno, cyn symud ymlaen i ddisgrifio sut y gweithredwyd y project.

2 TECHNOLEG CYFIEITHU

Mae technoleg cyfieithu'n cwmpasu amrywiaeth eang o ddulliau. Yn y bennod hon, byddaf yn canolbwyntio ar y ddwy dechnoleg sydd fwyaf perthnasol i'r bartneriaeth SMART: cofion cyfieithu a chyfieithu peirianyddol.

2.1 Cofion Cyfieithu

Yn eu hanfod, cronfeydd data yw cofion cyfieithu sy'n caniatáu i ddeunydd a gyfieithwyd yn flaenorol gael ei ail-ddefnyddio. Fe'u defnyddir yn gyffredinol yng nghyd-destun meddalwedd cyfieithu. Mae Ffigur 1 isod yn dangos cyfieithydd yn defnyddio cof cyfieithu i ôl-olygu dogfen yn Trados Studio.



Ffigur 1 – Cof cyfieithu yn cael ei ddefnyddio i olygu dogfen yn Trados Studio

Mae llif gwaith ôl-olygu sy'n defnyddio cofion cyfieithu fel arfer yn gweithio fel a ganlyn. Wrth i'r cyfieithydd weithio ar ddogfen, mae'r meddalwedd cyfieithu yn gwirio pob segment yn erbyn y cof cyfieithu. Os yw'r cof cyfieithu'n cynnwys segment tebyg, bydd yn tynnu sylw'r cyfieithydd at hyn ynghyd ag arwyddion gweledol sy'n cyfleu maint y gwaith golygu sy'n angenrheidiol. Yn ffigur 1 er enghraifft, mae segmentau wedi cael eu cyfateb i wahanol raddau. Mae rhai ohonynt yn gyfatebiadau llawn ac yn wyrdd tywyll. Efallai y bydd y cyfieithydd yn penderfynu gadael y segmentau hyn fel y maent. Mae segmentau eraill yn cyfateb yn rhannol, gyda chanran y cyfatebiad yn ymddangos mewn oren. Bydd y cyfieithydd fel arfer yn gwneud golygiadau i'r segmentau hyn gan

ddefnyddio set o arwyddion gweledol y meddalwedd fel canllaw. Yn y brif ffenestr olygu yn ffigur 1, mae Trados yn awgrymu bod y defnyddiwr yn dileu'r testun coch ac yn ychwanegu'r testun lliw gwyrddlas.

Mae corff mawr o ymchwil yn dangos mantais defnyddio cofion cyfieithu i gyfieithwyr, ac i gwmnïau cyfieithu [3]. Ond mae gan gofion cyfieithu oblygiadau pwysig y tu hwnt i hyn hefyd, gan eu bod yn gyfrwng storio delfrydol ar gyfer y testunau cyfochrog sy'n hanfodol ar gyfer cyfieithu peirianyddol a niwral.

2.2 Cyfieithu Peiriannol

Mae cyfieithu peirianyddol yn dechnoleg sy'n caniatáu i ddefnyddiwr gyfieithu testun yn awtomatig, hyd yn oed os nad yw'r system erioed wedi gweld y testun hwnnw o'r blaen. Mae'r ffordd y gwneir hyn yn amrywio'n sylweddol, ond y tri phrif batrwm yw cyfieithu peirianyddol seiliedig ar reolau, cyfieithu peirianyddol ystadegol a chyfieithu peirianyddol niwral. Cyfieithu peirianyddol seiliedig ar reolau oedd y paradeim cyntaf un ym maes cyfieithu peirianyddol. Mae'r dechneg yn gofyn am raglennu rheolau gramadegol penodol sy'n pennu sut bydd y peiriant yn cyfieithu o un iaith i'r llall. Y brif anfantais i'r mathau hyn o beiriannau yw eu diffyg hyblygrwydd, a'r angen am fuddsoddiad sylweddol mewn amser ar gyfer creu'r systemau [5].

Mae cyfieithu peirianyddol ystadegol (SMT) yn rhan o'r ail don ym maes technoleg cyfieithu. Yn hytrach na defnyddio rheolau sydd wedi'u rhaglennu â llaw, mae modelau SMT yn dysgu patrymau o gael eu hyfforddi ar ddata corpws cyfochrog. Yn y dull mwyaf llwyddiannus, cyfieithu peirianyddol ystadegol ar sail ymadroddion (PBSMT), caiff ymadroddion sy'n cyd-ddigwydd yn ystadegol yn y data eu sgorio a'u hechdynnu i dablau ymadroddion enfawr, y gellir eu hail-drefnu'n hwyrach i greu'r cyfieithiad terfynol. Cyfnod SMT oedd y cyntaf lle'r oedd cyfieithu peirianyddol yn cael ei ystyried yn ddigon da i'w ddefnyddio ar raddfa ddiwydiannol, gyda pheiriannau cyfieithu yn dod yn rhan o lif gwaith nifer o gwmnïau, a chyfieithwyr yn cael eu hyfforddi i ôl-olygu [6]. Un o anfoneision modelau SMT yw eu diffyg rhuglder cymharol (o'i gymharu â NMT), cymhlethdod pensaernïaeth y model ac ôl-troed mawr y tablau ymadroddion yn y cof. Modelau SMT oedd y n rhoi'r canlyniadau gorau mewn cyfieithu peirianyddol hyd nes i NMT gyrraedd yng nghanol y 2010au.

Mae cyfieithu peirianyddol niwral yn cynrychioli'r don ddiweddaraf mewn cyfieithu peirianyddol. Yn hytrach na defnyddio tablau ymadroddion, mae modelau NMT yn defnyddio rhwydweithiau niwral. Mae dwy ran i'r model – amgodiwr a datgodiwr – ac mae'r ddau yn cynnwys haenau. Mae'r segment sydd i'w gyfieithu yn cael ei brosesu gair wrth air gan yr amgodiwr, sy'n trosi pob gair yn fector. Mae fectorau yn ddilyniannau o rifau sy'n gallu cynrychioli gwybodaeth eithriadol o gyfoethog am bob gair mewn sawl dimensiwn, gan ganiatáu i'r model wneud rhagfynegiadau mwy effeithiol. Un o brif fanteision NMT yw ei fod yn gallu cynhyrchu cyfieithiadau sy'n fwy apelgar i werthuswyr dynol [7], [8]. Rhan o'r rheswm am hyn yw bod peiriannau cyfieithu niwral yn llawer llai tebygol o gynhyrchu gwallau gramadegol, gyda Bentivogli et al.[9] yn dangos eu bod yn cynhyrchu 19% yn llai o wallau geiriau, 17% yn llai o gamgymeriadau morffolegol a 50% yn llai o gamgymeriadau trefn geiriau. Fodd bynnag, mae'n ymddangos bod angen mwy o ddata hyfforddi ar fodelau NMT nag ar fodelau SMT er mwyn iddyn nhw berfformio'n effeithiol [10]. Mae tystiolaeth hefyd bod data hyfforddi o ansawdd gwael yn fwy niweidiol i beiriannau NMT [11].

Er mwyn gwneud dewisiadau o ran pa beiriannau cyfieithu y dylid eu defnyddio, mae angen dull dibynadwy o fesur ansawdd MT. Hyd at ddechrau'r unfed ganrif ar hugain, y dechneg fwyaf cyffredin oedd defnyddio gwerthuswyr dynol, er bod hon yn aml yn broses araf a chostus [11]. Ar hyn o bryd y metrig a ddefnyddir fwyaf yw BLEU [12]. Mae BLEU yn gweithio drwy gymharu allbwn peiriant cyfieithu – y rhagdybiaeth – gyda chyfieithiadau dynol o'r un data. Mae sgôr o sero yn cynrychioli rhagdybiaeth nad yw'n debyg o gwbl i'r cyfieithiad cyfeiriadol dynol, tra y byddai sgôr o 100 yn arwydd o ragdybiaeth sy'n cyfateb i'r cyfieithiad dynol yn union. Yn ôl Wołk a

Mazrek [13], gellir ystyried sgôr BLEU rhwng 15-30 yn gyfieithiad boddhaol, tra bo sgorau dros 50 yn cael eu hystyried yn rhai rhagorol. Nid yw sgoriau BLEU yn cael eu hystyried yn gymaradwy rhwng gwahanol ieithoedd, ac mae rhywfaint o dystiolaeth bod y metrig yn tueddu i danbriso systemau NMT o'i gymharu â rhai SMT [14].

2.3 Cyfieithu Peiriannol Mewn Cyd-destunau Lleiafrifol

Mae gwaith ar dechnoleg cyfieithu peirianyddol yng nghyd-destun Cymru wedi bod yn gymharol brin. Datblygodd Tyers a Donnelly [15] beiriant cyfieithu yn seiliedig ar reolau gan ddefnyddio fframwaith Apertium¹. Mae'r ymchwilwyr yn adrodd sgorau o 15.68 a 32.21 BLEU ar gyfer eu peiriant cyfieithu Cymraeg i Saesneg wrth gyfieithu sampl o Gofnodion y Cynulliad a'r Wicipedia Cymraeg, yn y drefn honno. Mae Jones ac Eisele [16] wedi disgrifio datblygiad peiriant SMT wedi'i hyfforddi ar gorpws cyfochrog a echdynnwyd o Gofnodion y Cynulliad. Hyfforddwyd y peiriant ar 510,813 segment cyfochrog ac fe gafodd sgôr o 36.17 BLEU ar gyfer y cyfeiriad iaith Saesneg i Gymraeg a 40.22 BLEU ar gyfer y cyfeiriad Cymraeg i Saesneg. Yn fwy diweddar, datblygodd Prys a Jones [4] gyfres o beiriannau cyfieithu parth-benodol ar gyfer y cwmni cyfieithu Cymen, gyda pheiriannau'n cael eu hyfforddi i gleientiaid ar sail eu data penodol nhw. Roedd y canlyniadau'n amrywio yn ôl maint data hyfforddi y cleientiaid, gan amrywio o 48.53 ar gyfer y cleient gyda'r set ddata leiaf i 59.06 ar gyfer y cleient gyda'r set ddata fwyaf. Disgrifir yr astudiaeth hon ymhellach yn adran 4 isod.

Y tu allan i'r cyd-destun Cymreig, mae gwaith a wnaed gan ymchwilwyr ar y pâr o ieithoedd en <> ga (Saesneg a Gwyddeleg) yn cynnig cymhariaeth ddiddorol. Mae Dowling et al. [17] yn disgrifio gwaith ar *Tapadóir*, peiriant cyfieithu en > ga SMT sydd â'r bwriad o hwyluso cyfieithu ar gyfer yr adran yn Llywodraeth Iwerddon sy'n gyfrifol am yr iaith Wyddeleg. Cafodd yr ymchwilwyr sgôr BLEU o 43.08 drwy gyfuno cofion cyfieithu parth benodol (*domain specific*) ag adnoddau eraill sydd ar gael yn gyhoeddus. Yn fwy diweddar, dywedodd yr un tîm fod ymgais gychwynnol i hyfforddi peiriant NMT gyda'r un set ddata wedi cael 6.40 pwynt BLEU yn llai na *Tapadóir* [10]. Eglurodd yr ymchwilwyr fod y canlyniad hwn yn gysylltiedig â natur gymharol gymhleth morffoleg y Wyddeleg ynghyd â phrinder y data hyfforddi, sy'n broblem gyffredin mewn cyd-destunau iaith leiafrifedig.

Ceisiodd astudiaeth ddilynol gan Defauw et al. [18] fynd i'r afael â'r problemau hyn gan ddefnyddio strategaeth casglu data mewn tri cham. Yn gyntaf, fe wnaeth yr ymchwilwyr graffu data cyfochrog ychwanegol o wefannau dwyieithog gan ddefnyddio'r teclyn *Scrapy*² ac yna eu cysoni â *Malign*. Yn ail, defnyddiodd yr ymchwilwyr gorpws cyfochrog newydd a oedd ar gael yn gyhoeddus fel rhan o'r ystorfa *ParaCrawl*³ [19]. Yn olaf, cafodd corpora synthetig eu creu drwy ôl-gyfieithu data corpws uniaith. Roedd angen glanhau'r holl adnoddau ychwanegol hyn gan ddefnyddio'r offeryn *Bicleaner* i gael gwared â brawddegau a oedd wedi'u cysoni'n wael neu a oedd yn peri problemau mewn ffyrdd eraill. Canfu'r ymchwilwyr fod model en > ga a hyfforddwyd gyda chymysgedd o ddata a grafwyd o'r we a data *ParaCrawl* wedi gwella 11 pwynt BLEU wrth gyfieithu set ddata parth benodol ac 8.7 BLEU wrth gyfieithu set ddata generig. Mae'r canfyddiadau hyn yn dangos, er bod angen llawer iawn o ddata ar gyfer hyfforddi modelau NMT effeithiol, y gallai crafu'r we a strategaethau ategu data eraill fod yn ateb ymarferol ar gyfer cyd-destunau iaith leiafrifedig a phrin ei hadnoddau.

¹Gellir mynd at yr peiriant drwy borwr yn [https://](https://www.apertium.org)

www.apertium.org. ²<https://scrapy.org/>

³<https://paracrawl.eu/>

3 CYMEN – CWMNI CYFIEITHU CYMRAEG ARLOESOL

Sefydlwyd Cymen yn 1987 fel rhan o don o gwmnïau cyfieithu preifat a ymddangosodd mewn ymateb i'r twf mewn galw am wasanaethau cyfieithu o'r Saesneg i'r Gymraeg ar y pryd [2]. Ers hynny, mae Cymen wedi tyfu i gyflogi 24 aelod o staff ynghyd â thros 20 cyfieithydd llawrydd allanol. Mae swyddfeydd y cwmni yng Nghaernarfon yng ngogledd Cymru, ond mae ei gwsmeriaid wedi'u lleoli ym mhob cwr o'r byd. Un patrwm sydd wedi dod i ddiffinio Cymen yw'r tueddiad i gymryd diddordeb agos mewn datblygiadau technolegol. Mae hyn wedi amlygu ei hun mewn cyfres o broiectau ymchwil ar y cyd â Phrifysgol Bangor rhwng 2000 a 2020. Cynhaliwyd y project cyntaf yn 2000 gyda'r nod o sefydlu defnydd cofion cyfieithu yn y cwmni. Roedd yr ail broiect, a gychwynnodd yn 2017, yn bartneriaeth dwy flynedd gyda'r nod o sefydlu'r defnydd o dechnoleg cyfieithu peirianyddol ystadegol (SMT). Amcan y project diweddaraf, a ddechreuodd yn 2019, oedd disodli'r peiriannau SMT hyn gyda rhai NMT.

Dylid pwysleisio nad yw'r math hwn o agwedd gadarnhaol tuag at dechnoleg o reidrwydd yn gyffredin yn y diwydiant cyfieithu. Yn ôl un arolwg diweddar, dim ond 22% o gyfieithwyr y DU a ddywedodd eu bod yn defnyddio technoleg cyfieithu peirianyddol, a dywedodd 70% o'r rhain mai *Google Translate* roedden nhw'n ei ddefnyddio [20]. Mae hyn yn codi cwestiwn pwysig – pam ddylai cwmni fel Cymen drafferthu creu llif gwaith cyfieithu peirianyddol cymharol gymhleth os yw'n bosibl defnyddio *Google Translate* neu *Microsoft Translator* yn Trados yn lle hynny? Mae sawl ateb i'r cwestiwn hwn. Y cyntaf yw bod defnyddio'r gwasanaethau hyn yn golygu anfon gwybodaeth o eiddo'r cwmni i ddwylo trydydd partion rhyngwladol. Gallai hyn arwain at oblygiadau cyfreithiol i gwmnïau sy'n gorfod delio â dogfennau sensitif neu gyfrinachol. Mae a wnelo'r ail ateb ag ansawdd a hyblygrwydd y gwasanaeth cyfieithu peirianyddol. Mae Google a Microsoft yn darparu peiriannau cyfieithu generig gyda'r bwriad o ddelio â'r amrywiaeth ehangaf posibl o destunau i'w cyfieithu. Drwy ddefnyddio eu harchifau eu hunain o gyfieithiadau blaenorol, fodd bynnag, gall cwmnïau cyfieithu gynhyrchu peiriannau cyfieithu sydd wedi'u teilwra i'r mathau o destun y maent yn fwyaf tebygol o'u cyfieithu.

4 PARTNERIAETHAU BLAENOROL CYMEN A PHRIFYSGOL BANGOR

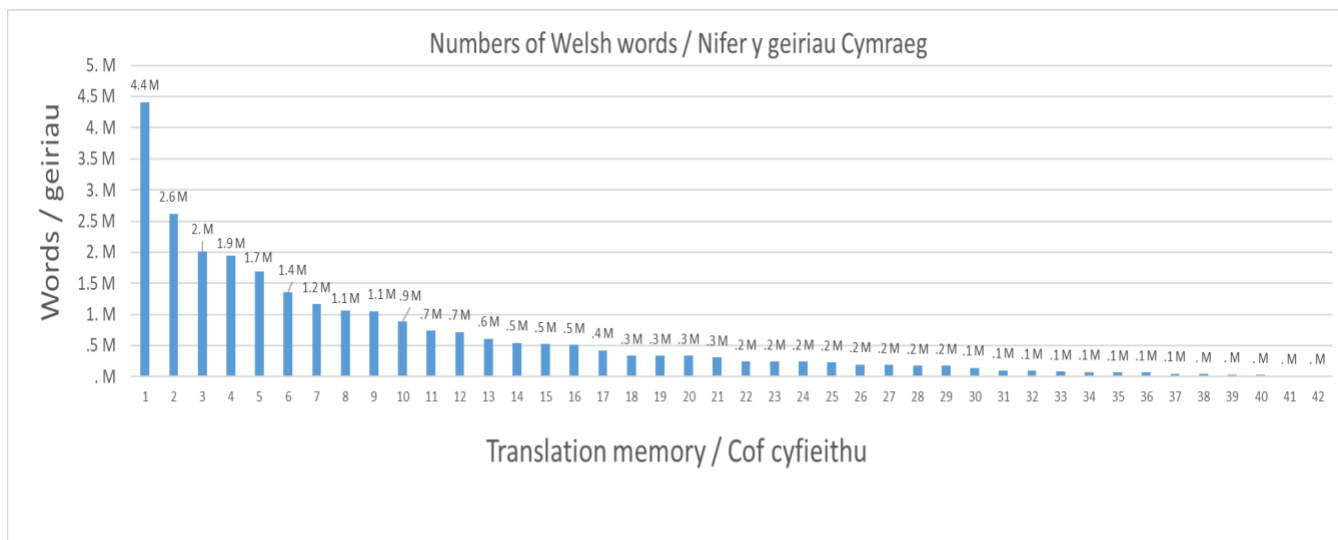
Rhoddwyd llifoedd gwaith cof cyfieithu presennol Cymen ar waith am y tro cyntaf yn ystod y project KTP cychwynnol yn 2000. Ers hynny, mae'r system wedi datblygu i fod yn syml ond yn effeithiol. Mae pob dogfen sydd i'w chyfieithu ar gyfer cleient yn cael ei chysylltu â chofion cyfieithu a rhestrau geirfa penodol sy'n berthnasol i'r cwsmer hwnnw. Mae'r ddogfen wedyn yn cael ei throsi'n broiect i'w gyfieithu yn Trados, gyda'r holl adnoddau cyfieithu perthnasol yn ymddangos yn awtomatig i'r cyfieithydd cyn gynted ag y bydd yn dechrau gweithio. Rheolir y system hon gyda chyfres o dempledi sy'n gosod ffurfweddiad y projectau a gynhyrchir ar gyfer pob cleient. Dros amser mae'r cwmni wedi adeiladu storfa fawr o gofion cyfieithu sy'n benodol i rai cleientiaid, yn ogystal â chofion cyfieithu mwy cyffredinol y gellir eu defnyddio gyda gwaith cleientiaid llai rheolaidd.

Dechreuodd ail broiect KTP yn 2017. Nod yr ail bartneriaeth hon oedd manteisio ar archif fawr y cwmni o gyfieithiadau blaenorol er mwyn hyfforddi peiriannau cyfieithu SMT a gwreiddio'r rhain yn llifoedd gwaith y cwmni. Disgrifir y broses hon yn fanwl yn Prys a Jones [4], felly dim ond crynodeb byr yw'r canlynol.

Cam cyntaf y project oedd dewis y ffordd orau o drosi archif y cwmni o gyfieithiadau blaenorol i ffurf briodol ar gyfer hyfforddi peiriannau cyfieithu. Penderfynwyd manteisio ar gyflenwad y cwmni o gofion cyfieithu at y diben hwn (gweler adran 2.1). Mae cofion cyfieithu yn cael eu storio ar ffurf ffeiliau TMX, sy'n defnyddio iaith farcio XML i strwythuro'r data. Ysgrifennwyd piblinell cod i awtomeiddio'r broses o dynnu pob segment gyfochrog o gofion cyfieithu Cymen a chreu corpws hyfforddi yn awtomatig. Ychwanegwyd cod pellach er mwyn glanhau a pharatoi'r data ar gyfer y broses hyfforddi. Ymysg pethau eraill, mae hyn yn cynnwys tynnu marciau XML o'r

segmentau yn ogystal â dileu unrhyw segmentau lle mae'r cyfieithiad gryn dipyn yn hirach neu'n fyrrach na'r gwreiddiol. Y canlyniad yw corpws cyfochrog sy'n lân ac yn barod ar gyfer y broses hyfforddi.

Hwyluswyd yr hyfforddiant gan ein gallu i fanteisio ar waith blaenorol gan Uned Technoleg Iaith Prifysgol Bangor, rhaglen hyfforddi SMT sy'n defnyddio'r rhaglen *Moses SMT*. Mae'r pecyn hwn ar gael yn rhad ac am ddim ar dudalen Github yr uned⁴. Ymysg yr ychwanegiadau a wnaed yn ystod y project roedd camau ar gyfer tiwnio a gwerthuso data yn awtomatig gan ddefnyddio BLEU. Yn dilyn hyn, datblygwyd ap i gyflwyno'r peiriannau i'r amgylchedd cyfieithu a ddefnyddir gan gyfieithwyr y cwmni. Mae gwneuthurwyr Trados wedi galluogi datblygu ategion cod agored gan ddefnyddio templedi cod a phrojectau parod sydd ar gael ar eu tudalen Github⁵. Addaswyd un o'r ategion rhagosodedig⁶ hyn i greu ein hategyn personol ein hun.



Ffigur 1 – Nifer y miliynau o eiriau yng nghof cyfieithu Cymen (2020)

Dechreuodd y cwmni ddefnyddio'r peiriannau yn raddol, gyda'r rheolwyr yn eu profi i ddechrau cyn caniatáu eu defnyddio'n fwy eang. Roedd unrhyw beiriannau cyfieithu a oedd yn ymddangos yn aneffeithiol yn cael eu gollwng ar unwaith. Yn y pen draw, dim ond dau beiriant cyfieithu a oroesodd y broses werthuso hon, y rhai a hyfforddwyd ar gofion cyfieithu 1 a 2 (gweler ffigur 2 uchod). Cafwyd bod y peiriant a hyfforddwyd ar TM 1 yn arbennig o ddefnyddiol gan y cyfieithwyr. Roedd llwyddiant y peiriant cyfieithu penodol hwn yn bennaf oherwydd ei fod yn bodloni dau o brif feini prawf cyfieithu peirianyddol – data arbennig o berthnasol ynghyd â set hyfforddi fawr [6]. Cafodd y peiriant hwn ei osod i ymddangos yn awtomatig ar gyfer cyfieithwyr ym mhob project cyfieithu ar gyfer y cleient hwn, ac roedd yr adborth gan gyfieithwyr yn gadarnhaol. Nid oedd ansawdd y peiriant cyfieithu parth cyffredinol (hynny yw, peiriant wedi'i hyfforddi ar yr holl ddata sydd ar gael) yn cael ei ystyried yn ddigon da i'w ddefnyddio yn y rhan fwyaf o achosion ac aeth heb ei ddefnyddio ar y cyfan.

⁴<https://github.com/PorthTechnolegaulaith/moses-smt>

⁵<https://github.com/sdl/Sdl-Community>

⁶<https://github.com/OpenNMT/Plugins>

5 PARTNERIAETH SMART A NMT

Yn dilyn yr ail broiect KTP, dywedwyd wrthym fod partneriaeth newydd i uwchraddio'r peiriannau cyfieithu wedi'i chymeradwyo.

Cam cyntaf y project oedd dewis pecyn meddalwedd ar gyfer hyfforddi modelau NMT. Ar ôl cyfnod byr o ymchwili, dewiswyd Marian NMT gan ei fod yn cynnig llawer o nodweddion hyfforddi arloesol ynghyd â dogfennaeth glir a hygyrch. Yn dilyn hyn, treuliodwyd amser yn casglu data hyfforddi ychwanegol. Un bwriad a oedd gennym ar gyfer y project hwn oedd ceisio datblygu peiriant parth cyffredinol a fyddai'n gallu trin data unrhyw gleient i lefel foddhaol. I'r perwyl hwnnw fe wnaethom ni geisio casglu cymaint o ddata Cymraeg <> Saesneg â phosibl o wahanol ffynonellau.

Tabl 1: Data hyfforddiant ar gyfer peiriant parth cyffredinol Cymen

Data	Nifer y geiriau (Cymraeg)
Cofion cyfieithu Cymen	89 miliwn
Cofnodion y Cynulliad Cenedlaethol + Deddfwriaeth + Meddalwedd	18 miliwn
Cofion cyfieithu Llywodraeth Cymru	2 miliwn
Rhestrau termau Llywodraeth Cymru	0.4 miliwn
Cyfanswm	109.4 miliwn
Cyfanswm ar ôl glanhau	86 miliwn

Fel y gwelir yn nhabl 1 uchod, casglwyd data ychwanegol o amrywiaeth o ffynonellau i ategu setiau data'r cwmni ei hun. Y brif ffynhonnell oedd archif fewnol Cymen o gofion cyfieithu, sydd wedi tyfu i 89 miliwn o eiriau Cymraeg ers y project blaenorol. Daeth rhan o'r twf hwn o hen ffeiliau TMX (yn ogystal â fformatau ffeiliau eraill) a ganfuwyd yn archifau Cymen. Casglwyd hyd yn oed mwy o ddata yng nghwrs gwaith cyfieithu dyddiol y cwmni, gydag oddeutu 1 miliwn o eiriau'n cael eu hychwanegu at atgofion cyfieithu Cymen ym mhob mis yn ystod 2020.

Yr ail ffynhonnell oedd <http://techiaith.cymru>. Mae'r wefan hon yn darparu tri chorpws cyfochrog fel rhan o addasiad yr Uned Technoleg Iaith o'r rhaglen Moses SMT: corpwys a echdynnwyd o drafodion y Cynulliad Cenedlaethol, corpwys o ddeddfwriaeth ddwyieithog Llywodraeth y DU, a chorpws o gyfieithiadau meddalwedd. Yn olaf, mae Llywodraeth Cymru wedi dechrau cynnig ffeiliau TMX ar eu gwefan yn ddiweddar at ddefnydd y cyhoedd yn <https://llyw.cymru/bydtermcymru> <about:blank>. Mae'r storfa hon yn cynnig amrywiaeth eang o ffeiliau TMX sy'n ymwneud â meysydd fel ieuchyd, addysg a deddfwriaeth. Ar hyn o bryd (Tachwedd 2020) mae'r ffeiliau'n cynnwys tua 2 filiwn o eiriau Cymraeg i gyd, ond mae'n ymddangos bod y ddarpariaeth yn tyfu'n gyflym. Mae'r wefan hefyd yn cynnig nifer o restr termau sydd wedi'u safoni gan Lywodraeth Cymru ac sydd ar gael mewn fformat ffeil Excel sy'n hawdd i'w trosi'n gorpora cyfochrog.

Ceir corpora cod agored ychwanegol ar gyfer y Gymraeg a'r Saesneg yn ystorfa Opus⁷, ond fe wnaethon ni benderfynu peidio â defnyddio'r rhain oherwydd ansawdd eithriadol o wael y data.

Ar ôl cyfuno'r corpora hyn a chael gwared ar segmentau dyblyg, ceir corpwys cymharol sylweddol o 109 miliwn o eiriau Cymraeg. Ar ôl y broses glanhau (fel sy'n cael ei ddisgrifio yn adran 4), mae'r corpwys yn cael ei leihau i oddeutu 86 miliwn o eiriau. O ystyried maint cymharol fach y corpora Cymraeg <> Saesneg sydd ar gael i'r cyhoedd (gweler tabl 1), ymddengys bod sefyllfa Cymen yn arbennig o addawol o ran hyfforddi peiriannau NMT.

⁷<http://opus.nlpl.eu/>

5.1 Ffurfweddu'r Model NMT

Elfen arbennig o bwysig wrth hyfforddi modelau NMT yw tiwnio'r hyperbaramedrau, sef yr elfennau hynny o'r model y gall yr ymchwilydd eu ffurfweddu ei hun. Mae'r rhain yn cynnwys dulliau cysoni fel gollyngiad [21] a llyfnhau esbonyddol, sy'n helpu i atal y model rhag gorfffitio i'r data hyfforddiant. Gall ymchwilydd hefyd ddewis elfennau o bensaerniaeth y system, gyda Marian yn caniatáu dewis rhwng model RNN sylfaenol, model seq2seq, a model trawsnewidydd [22] yn ogystal â dimensiynau'r datgodiwr a'r amgodiwr. Mae nodweddion eraill fel normaleiddio haenau [23] yn caniatáu lleihau amser hyfforddi'r model.

Dros gyfnod o sawl arbrawf, fe wnaethon ni brofi sawl ffurfweddiad gwahanol, gan ddefnyddio'r sgorau BLEU fel canllaw ar gyfer effeithiolrwydd unrhyw newidiadau. Mae rhai o'r hyperbaramedrau ar gyfer y model gorau a geir i'w gweld yn nhabl 2 isod.

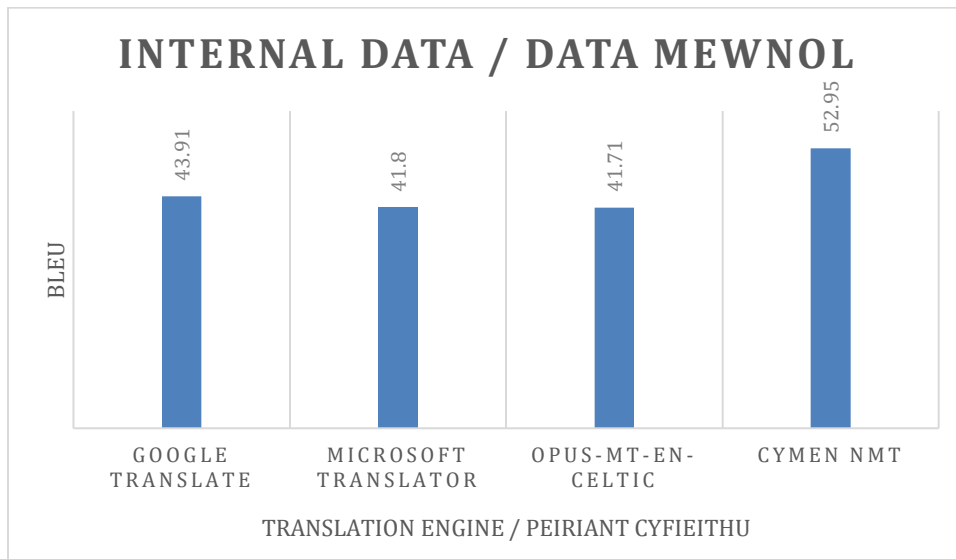
Tabl 2: Ffurfweddiad y model parth cyffredinol Cymraeg <> Saesneg

Nodwedd	Gwerth
Pensaerniaeth y model	Trawsnewidydd
Gollyngiad y ffynhonnell	0.1
Gollyngiad y targed	0.1
Dyfinder yr amgodiwr	4
Dyfinder y datgodiwr	4
Normaleiddio haenau	-
Llyfnhau esbonyddol	0.0001
Alinio dan arweiniad	-

Yn ystod oes y project, cyflawnwyd cynnydd o tua 5 pwynt BLUE drwy wneud newidiadau i ffurfweddiad y model. Mae hyn yn dangos pwysigrwydd peidio â gadael y model yn ei ragosodiadau gwreiddiol, ond gwneud cymaint o welliannau â phosibl yn lle hynny.

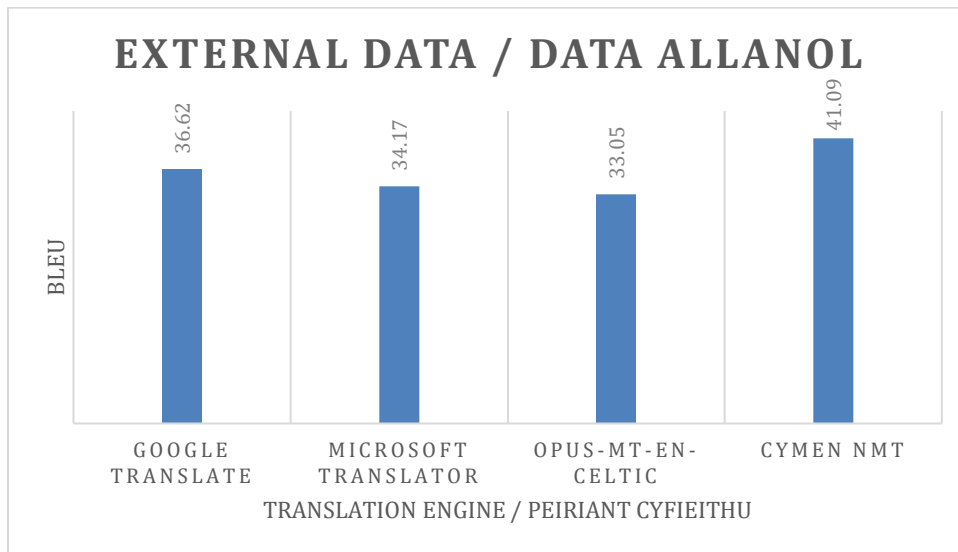
5.2 Ansawdd y Peiriant Cyfieithu a Chymhariaeth

Mae Ffigur 2 yn dangos y canlyniadau ar gyfer y peiriant gorau o'i gymharu â'r 3 pheiriant cyfieithu en > cy sydd ar gael i'r cyhoedd ac sy'n hysbys i ni. Mae Google Translate a Microsoft Translator yn wasanaethau adnabyddus sy'n seiliedig ar API a phorwr a ddarperir gan gwmnïau rhyngwladol, ac mae OPUS-MT-EN-CELTIC yn fodel newidydd cod agored sydd ar gael i'w lwytho i lawr oddi ar wefan Hugging Face. Mae pob un o'r pedwar yn beiriant cyfieithu niwral.



Ffigur 2 – Ansawdd y peiriant wrth gyfieithu data mewnlol Cymen

Daw'n amlwg ar unwaith bod peiriant NMT Cymen yn cyfieithu data mewnlol y cwmni i safon llawer uwch na'r gwasanaethau eraill, gan berfformio tua 9 BLEU yn well ar y dasg hon na'r cystadleuydd agosaf, *Google Translate*. Yr eglurhad mwyaf tebygol am hyn yw natur data hyfforddi Cymen. Mae gan Cymen fynediad at lawer iawn o ddata hynod berthnasol yn mynd yn ôl 20 mlynedd, gan gynnwys cyfieithiadau blaenorol ar gyfer cleientiaid sydd wedi cyflwyno gwaith yn gyson dros y cyfnod hwnnw. Nid yw'n syndod felly bod peiriant Cymen yn perfformio'n well na'r gystadleuaeth yma, gan nad oes gan y sefydliadau eraill fynediad at ddata'r cwmni.



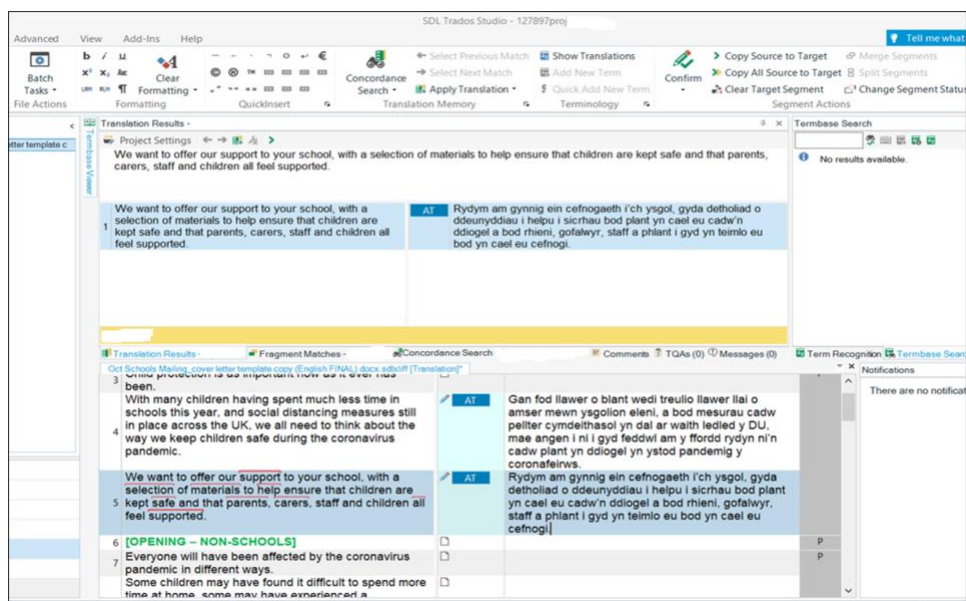
Ffigur 3 – Ansawdd y peiriant wrth gyfieithu data allanol Cymen

Dengys Ffigur 3 ansawdd cyfieithiadau'r peiriannau cyfieithu wrth gyfieithu data allanol nad yw'n perthyn i Cymen. Cynhyrchwyd y set brawf hon o gyfieithiadau blaenorol a wnaed gan uned gyfieithu Prifysgol Bangor. Fel set ddata sy'n perthyn i sefydliad sydd â'i arddull tŷ ei hun, gellir dadlau y gallai hyn fod yn brawf teg o ansawdd y peiriannau hyn ar 'ddata niwtral'. Mae peiriant NMT Cymen unwaith eto'n perfformio ar y blaen i'r peiriannau eraill yma, er bod y gwahaniaeth ychydig yn llai – tua 5 pwynt BLEU y tro hwn.

Beth sy'n egluro perfformiad rhagorol peiriant NMT Cymen yn yr achos yma? Un posibilrwydd yw bod gan Cymen fwy o ddata o ansawdd uchel ar gyfer y pâr ieithyddol yma. Daw'r rhan fwyaf o'r data hyfforddi o gofion cyfieithu Cymen ei hun, sy'n cynnwys miliynau o eiriau o gyfieithiadau proffesiynol, sydd wedi'u prawfddarllen. Gellid dadlau bod meddu ar archif o'r fath yn fantais sylweddol i'r cwmni dros sefydliadau eraill.

Mae hefyd yn werth nodi'r gwahaniaethau rhwng y peiriannau cyfieithu eraill. Ymddengys mai *Google Translate* oedd yr ail fwyaf llwyddiannus yn y ddau arbrawf, gan sgorio tua 3 phwynt yn uwch na *Microsoft Translator*. Mae OPUS-MT-EN-CELTIC a *Microsoft Translator* yn gymharol agos – gydag oddeutu 1 pwynt BLEU rhyngddynt yn y ddau achos. Mae hyn yn ganlyniad trawiadol i'r peiriant *Hugging Face*, gan ei bod yn debyg bod y gwneuthurwyr wedi'u cyfyngu i ddata hyfforddi sydd ar gael i'r cyhoedd⁸.

⁸Mae tudalen y model yn rhestru Opus, Techiaith a BT (*back translations*, sef ôl-gyfieithiadau) fel setiau data a ddefnyddiwyd.



Figur 5 - Peiriant en > cy Marian Cymen yn rhedeg yn Trados

Ar ôl optimeiddio'r peiriannau, y cam nesaf oedd galluogi cyfieithwyr i'w defnyddio. I wneud hyn, fe wnaethom addasu'r ategyn a ddatblygwyd yn y project blaenorol i ganiatáu i beiriannau cyfieithu Marian gael eu defnyddio yn Trados 2019. Fel yn y project blaenorol, bu i ni gyflwyno'r peiriannau newydd yn raddol, ond daeth yn amlwg yn sydyn fod agweddau'r cyfieithwyr tuag at NMT yn llawer mwy cadarnhaol na'r patrwm blaenorol. Dros amser penderfynodd y cwmni ddefnyddio'r dechnoleg ym mhob un project yn ddiofyn, sy'n golygu bod cyfieithu peirianyddol bellach yn rhan annatod o lifoedd gwaith y cwmni.

5.3 Adborth gan y Cyfieithwyr

Er mwyn dysgu mwy am brofiadau'r cyfieithwyr o weithio gyda'r peiriannau cyfieithu, fe wnaethon ni benderfynu cynnal cyfres o sgysiau grŵp ffocws a oedd yn cynnwys y swyddog ymchwil (fi), rheolwyr y cwmni a rhai o'r cyfieithwyr. Yn ystod y cyfarfodydd hyn, roedd cyfieithwyr yn gallu trafod eu profiadau cadarnhaol a negyddol o weithio gyda'r peiriannau cyfieithu. Roedd yr adborth yn gadarnhaol iawn ar y cyfan, yn enwedig wrth i'r project fynd rhagddo ac wrth i'r peiriannau cyfieithu gael eu hoptimeiddio. Roedd nifer o gyfieithwyr yn honni bod y peiriannau cyfieithu yn caniatáu iddynt weithio'n gynt ac yn rhoi mwy o amser iddynt brawf ddarllen. Honnodd rai o'r cyfieithwyr fod defnyddio'r peiriannau yn caniatáu iddynt fwy na dyblu nifer y geiriau roedd modd iddynt eu cyfieithu mewn diwrnod⁹. Yn ogystal â helpu i gyflymu'r gwaith, gallai'r peiriannau cyfieithu hefyd fod o gymorth seicolegol i gyfieithwyr sy'n gweithio ar broiectau arbennig o hir, oherwydd bod llai o ymdrech yn gysylltiedig â nhw.

Ymhlith y sylwadau negyddol mwyaf cyffredin roedd tueddiad i'r peiriannau cyfieithu rhifau ac enwau priod yn anghywir. Ar ôl i ni gael gwybod am y problem, roedd modd i ni ddatblygu ateb syml drwy wneud newidiadau i'r

⁹Er bod hyn wrth gwrs yn dystiolaeth anecdotaidd ar hyn o bryd, rydym yn gobeithio mesur y cynnydd canfyddedig mewn cynhyrchiant mewn modd empirig yn fuan.

adnodd CAT a ddefnyddir gan ein cyfieithwyr. Mae gan y rhan fwyaf o raglenni cyfieithu nodweddion QA (*Quality Assurance*) erbyn hyn sy'n caniatáu i gyfieithwyr ddilysu dogfen er mwyn sicrhau bod rhifau'n aros yr un fath yn y segmentau ffynhonnell a tharged, ac i atal cyfieithu termau pwysig yn anghywir. Mae'r holl broiectau bellach yn cael eu creu gyda'r gosodiadau hyn ar waith, gan ddiogelu yn erbyn y broblem.

Problem arall a nodwyd yn y grŵp ffocws oedd anallu'r peiriant i ddelio â marciau XML wrth gyfieithu. Mae tagiau XML yn aml yn cynnwys gwybodaeth hanfodol ar gyfer ail-greu fformat dogfennau, felly mae sicrhau bod tagiau'n cael eu gosod yn gywir o gwmpas y geiriau perthnasol yn y segment targed yn rhan hollbwysig (ond rhwystredig) o waith dyddiol cyfieithydd. Mae peiriannau SMT yn gallu delio â hyn yn gymharol dda, ond mae'n dal i fod yn broblem anodd ei datrys yn achos NMT. Mae'r strategaethau a awgrymwyd wedi cynnwys chwistrellu tagiau i gorpora cyfochrog ar gyfer ymestyn data [24], ond am nawr rydyn ni'n datrys y broblem drwy dynnu marciau XML yn yr ategyn yn awtomatig ac yna gadael i gyfieithwyr ychwanegu'r tagiau eto â llaw.

O bryd i'w gilydd, roedd cyfieithwyr wedi crybwyll gwallau a oedd yn ymwneud yn fwy ag amrywiad cywair naturiol y data hyfforddi nag unrhyw broblemau gyda'r peiriannau cyfieithu eu hunain. Fel cwmni cyfieithu, mae gan Cymen ei arddull ei hun y glynir ati fel arfer mewn achosion nad ydynt yn mynd yn groes i reolau penodol cleientiaid. Mewn rhai achosion, roedd y cyfieithwyr yn gwrthwynebu cyfieithiadau a gynhyrchwyd gan y peiriant cyfieithu a oedd yn torri naill ai arddull y tŷ neu ddewis y cyfieithydd ei hun. Un enghraifft o hyn oedd cyfieithu'r gair Saesneg '*whether*' i'r Gymraeg. Cyfieithir hyn yn amrywiol yn y data hyfforddi fel p'un ai a yw, p'un a yw, p'un ai yw, a mwy. Ar ôl ymgynghori â'r cyfieithwyr, fe wnaethon ni benderfynu ychwanegu rheolau ôl-olygu awtomatig at biblinell y peiriannau cyfieithu, gan sicrhau bod y ffurf a ffafrir - p'un a yw - yn cael ei defnyddio bob amser yn lle'r ffurfiau eraill. Yn dilyn llwyddiant y dull hwn, rydym yn awr yn cadw ffeil y gall cyfieithwyr ei defnyddio i ychwanegu rheolau tebyg eu hunain yn uniongyrchol. Efallai nad yw'r mathau hyn o fesurau'n dechnolegol soffistigedig, ond teimlwn eu bod yn bwysig yn yr ystyr eu bod yn cynyddu ymdeimlad y cyfieithwyr o berchnogaeth dros y dechnoleg, yn ogystal â gwella ansawdd yr allbwn yn fwy cyffredinol.

6 Y CAMAU NESAF

Mae sawl ffordd bosibl o wella peiriannau cyfieithu Cymen ymhellach. Y ffordd symlaf o wneud hyn fyddai casglu data cyfieithu ychwanegol. I ryw raddau bydd hyn yn digwydd yn naturiol yn achos Cymen gan fod y cwmni ar hyn o bryd yn ychwanegu oddeutu 1 miliwn o eiriau'r mis at ei gofion cyfieithu. Byddai hefyd yn ddoeth cynnal ymdrech, ynghyd â rhanddeiliaid eraill, i roi pwysau ar Lywodraeth Cymru i ryddhau rhagor o adnoddau cyfieithu dwyieithog (ffeiliau TMX, rhestrau termau a chorpora cyfochrog). Dull arall o gasglu data ychwanegol yw defnyddio gwe-ymlusgwyr. Mae Bitextor [25], er enghraifft, yn adnodd cod agored sy'n gallu ymlusgo drwy destun cyfochrog o wefannau dwyieithog, alinio segmentau i greu corpws cyfochrog gan ddefnyddio Marian, ac yna glanhau'r corpws a grëwyd gyda'r offeryn Bicleaner. Mae'r adnoddau hyn eisoes wedi cael eu defnyddio gyda thipyn o lwyddiant i gynyddu maint y data hyfforddi sydd ar gael ar gyfer y Wyddeleg (gweler adran 2.3).

Dull addawol arall yw creu corpora synthetig. Mae'r dechneg hon yn seiliedig ar y ffaith ei bod fel arfer yn llawer haws dod o hyd i ddata uniaith na data dwyieithog, yn enwedig ar gyfer deunydd parth penodol. Yn gyntaf, defnyddir peiriant cyfieithu sydd wedi'i hyfforddi'n barod i gyfieithu set ddata uniaith i greu ochr ffynhonnell synthetig, mewn proses a elwir yn ôl-gyfieithu. Gellir cyfuno'r corpws synthetig a geir wedyn â chorpws cyfochrog arferol. Mae ymchwilwyr wedi nodi gwelliannau sylweddol o'r dechneg hon, gyda Sennrich et al.[26] yn disgrifio cynnydd o rhwng 1 a 4 pwynt BLEU ar gyfer eu modelau nhw. Gellir cyfuno'r dull hwn hefyd â'r adnoddau Bitextor a Bicleaner a drafodwyd uchod i fireinio'r data.

7 CASGLIAD

Mae'r bennod hon wedi disgrifio project syml gyda manteision pellgyrhaeddol i gwmni cyfieithu Cymraeg. Yn wahanol i'r dechnoleg SMT flaenorol, cafodd y peiriannau cyfieithu NMT eu croesawu'n frwd gan gyfieithwyr y cwmni - yn bennaf oherwydd eu rhuglder uwch - ac maent bellach yn rhan annatod o lifoedd gwaith cyfieithu Cymen. Rydym hefyd wedi dangos manteision defnyddio peiriannau cyfieithu sydd wedi'u hyfforddi ar archif hen gyfieithiadau Cymen dros wasanaethau a ddarperir gan drydydd partion fel Google a Microsoft. Nid yn unig y mae peiriannau pwrpasol Cymen yn wrthrychol well am gyfieithu data perthnasol (gweler adran 5.2), ond gallent hefyd fod yn opsiwn gwell ar gyfer cyfieithu dogfennau sensitif drwy beiriant, ac maent yn sicr yn cynnig mwy o hyblygrwydd wrth ganiatáu i'r cwmni ymateb i anghenion cyfieithwyr.

CYDNABYDDIAETH

Cafodd y project hwn ei ariannu'n rhannol gan Lywodraeth Cymru drwy'r cynllun partneriaeth SMART. Hoffwn ddiolch i Samantha Williams a Natalie Crawley o Lywodraeth Cymru am eu cefnogaeth ddifflino dros y flwyddyn ddiwethaf. Diolch hefyd i staff Cymen Cyf am eu hyder a'u ffyddlondeb yn ystod y project hwn a'r project KTP blaenorol, ac am gadw meddwl agored drwyddo draw. Yn olaf, hoffwn ddiolch i Dewi Bryn Jones, Prifysgol Bangor, am ei waith caled a'i amynedd fel goruchwyliwr fy mhroject.

CANOLWYR

- [1] Llywodraeth Cymru. 2019. 2050: Miliwn o siaradwyr Cymrae. Adroddiad Blynyddol. Cyrchwyd Rhagfyr 02, 2020 o <https://gov.wales/sites/default/files/publications/2019-03/cymraeg-2050-a-millionwelsh-speakers-annual-report-2017-18.pdf>
- [2] Tegau Andrews. 2015. Cyd-destun gwleidyddol a chymdeithasol cyfieithu yn y Gymru gyfoes. Yn: Prys, D & Trefor, R. (Eds). Ysgrifau a Chanllawiau Cyfieithu. [Ar-lein]. Coleg Cymraeg Cenedlaethol, Caerfyrddin. Cyrchwyd Hydref 27, 2020 o <https://llyfrgell.porth.ac.uk/media/ysgrifau-a-chanllawiau-cyfieithu-delythprys-arobab-trefor-goln>
- [3] Ben Screen. 2018. Defnyddio Cyfieithu Awtomatig a Chof Cyfieithu wrth gyfieithu o'r Saesneg i'r Gymraeg: Astudiaeth ystadegol o ymdrech, cynhyrchedd ac ansawdd gan ddefnyddio data Cofnodwyr Trawiadau Bysell a Thracio Llygaid. Thesis PhD, Adran y Gymraeg, Prifysgol Caerdydd.
- [4] Myfyr Prys a Dewi Bryn Jones. 2019. Embedding English to Welsh MT in a Private Company. Yn Proceedings of the Celtic Language Technology Workshop. European Association for Machine Translation, Duly, 41-47.
- [5] Antonio Lagarda, Vicent Alabau, Francisco Casacuberta, Roberto Silva, ac Enrique Díaz-de-Liaño. 2009. Statistical post-editing of a rule-based machine translation system. Yn Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Association for Computational Linguistics, Boulder, Colorado, USA, 217-220.
- [6] Phillip Koehn. 2009. Statistical machine translation. Cambridge University Press.
- [7] Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, ac Andy Way. 2017. Is neural machine translation the new state of the art? The Prague Bulletin of Mathematical Linguistics, 108, 1 (Mehefin 2017), 109-120.
- [8] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névél, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor a Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. Yn Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. Association for Computational Linguistics, Berlin, Yr Almaen, 131-198.
- [9] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, a Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. Yn Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, 257-267.
- [10] Meghan Dowling, Teresa Lynn, Alberto Poncelas, ac Andy Way. 2018. SMT versus NMT: Preliminary Comparisons for Irish. Yn Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018). Association for Machine Translation in the Americas, Boston, USA, 12-20.
- [11] Phillip Koehn. 2020. Neural machine translation. Cambridge University Press.
- [12] Kishore Papineni, Salim Roukos, Todd Ward, a Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. Yn Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, Philadelphia, USA, 311-318.
- [13] Krzysztof Wołk a Krzysztof Marasek. 2015. Neural-based machine translation for medical text domain based on European medicines agency

leaflet texts. *Procedia Computer Science*, 64, (Hydref 2015), 2-9.

- [14] Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'dowd, ac Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32, 3 (Medi 2018), 217-235.
- [15] Francis Tyers a Kevin Donnelly. 2009. apertium-cy - a collaboratively-developed free RBMT system for Welsh to English. *The Prague Bulletin of Mathematical Linguistics*, 91 (Ionawr 2009), 57-66.
- [16] Dafydd Jones ac Andreas Eisele. 2006. Phrase-based statistical machine translation between English and Welsh. Yn *Strategies for developing machine translation for minority languages* (5th SALTML workshop on Minority Languages), LREC-2006. LREC, Genoa, Yr Eidal, 75-78.
- [17] Meghan Dowling, Lauren Cassidy, Eimear Maguire, Teresa Lynn, Ankit Srivastava a John Judge. 2015. Tapadóir: Developing a statistical machine translation engine and associated resources for Irish. Yn *Proceedings of the The Fourth LRL Workshop: "Language Technologies in support of Less-Resourced Languages"*. LRL, Poznan, Gwlad Pwyl.
- [18] Arne Defauw, Sara Szoc, Tom Vanallemeersch, Anna Bardadym, Joris Brabers, Frederic Everaert, Kim Scholte, Koen Van Winckel a Joachim Van den Bogaert. 2019. Developing a neural machine translation system for Irish. Yn *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. European Association for Machine Translation, Duly, 32-38.
- [19] Marta Bañón, Pinzhen (Patrick) Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Espla-Gomis, Mikel L. Forcada, et al. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. Yn *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Ar-lein, 4555-4567.
- [20] European Commission Representation in the UK, Chartered Institute of Linguists and the Institute of Translation and Interpreting. 2017. 2016 UK Translator Survey - Final Report. Cychwyd Medi 28, 2020 o https://ec.europa.eu/unitedkingdom/sites/unitedkingdom/files/ukts2016-final-report-web-_18_may_2017.pdf
- [21] Gal Yarin a Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 29, (Rhagfyr 2016), 1019-1027.
- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, ac Illia Polosukhin. 2017. Attention is all you need. Yn *Advances in neural information processing systems*. NeuralIPS, Long Beach, CA, UDA, 5998-6008.
- [23] Jimmy Ba, Jamie Kiros a Geoffrey Hinton. 2016. Layer normalization. Yn *Advances in NIPS 2016 Deep Learning Symposium*. NeuralIPS, Barcelona, Sbaen, 1-14.
- [24] Greg Hanneman a Georgiana Dinu. 2020. How Should Markup Tags Be Translated?. Yn *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, Ar-lein, 1160-1173.
- [25] Miquel Espla-Gomis. 2009. Bitextor, a free/open-source software to harvest translation memories from multilingual websites. Yn *Proceedings of MT Summit XII, Association for Machine Translation in the Americas*. Association for Machine Translation in the Americas, Ottawa, Canada.
- [26] Rico Sennrich, Barry Haddow, ac Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. Yn *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Yr Almaen<lt8/>, 86-96.