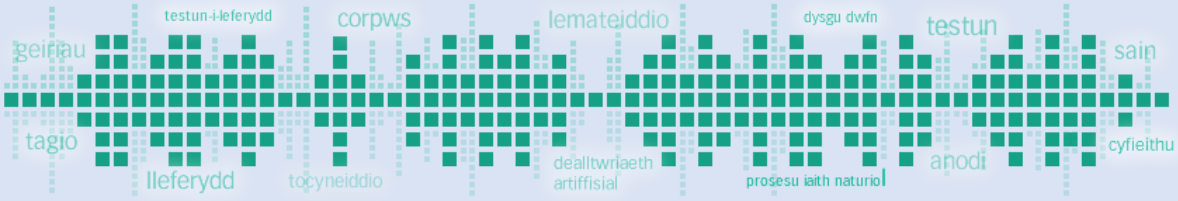
Iaith a Thechnoleg yng Nghymru: Cyfrol II

Watkins, Gareth; Prys, Delyth; Prys, Gruff; Jones, Dewi; Ghazzali, Stefano; Vangberg, Preben; Farhat, Leena; Cooper, Sarah; Williams, Meinir; Gruffydd, Ianto; Jouitteau , Mélanie; Grobol, Loïc; Morris , Jonathan; Ezeani , Ignatius; Young , Katharine; Davies, Lynne; El-Haj , Mahmoud; Knight , Dawn; Jarvis , Colin; Barnes, Emily

Cyhoeddwyd: 01/11/2024



Iaith a Thechnoleg yng Nghymru: Cyfrol II

Golygydd: Gareth Watkins

Cyhoeddwyd yr e-lyfr hwn gyntaf yn 2024 gan

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

www.bangor.ac.uk/cy

Rhif Llyfr Rhyngwladol (e-lyfr):

ISBN: 978-1 84220-208-1.

Cymorth dylunio a phrawfddarllen gan yr Athro Delyth Prys a Stefano Ghazzali. Mae'r llyfr hwn ar gael hefyd yn Saesneg dan y teitl *Language and Technology in Wales Volume 2*, rhif ISBN 978-1 84220-207-4.

Iaith a Thechnoleg yng Nghymru: Cyfrol II

Golygydd:
Gareth Watkins

Cyfranwyr:
Mélanie Jouitteau UNIVERSITE DE PAU ET DES PAYS DE L'ADOUR, A UNIVERSITÉ BORDEAUX-MONTAIGNE

Loïc Grobol UNIVERSITÉ PARIS NANTERRE

Jonathan Morris PRIFYSGOL CAERDYDD

Ignatius Ezeani PRIFYSGOL CAERHIRFRYN

Ianto Gruffydd PRIFYSGOL BANGOR

Katharine Young PRIFYSGOL CAERDYDD

Lynne Davies PRIFYSGOL CAERDYDD

Mahmoud El-Haj PRIFYSGOL CAERHIRFRYN

Dawn Knight PRIFYSGOL CAERDYDD

Preben Vangberg PRIFYSGOL BANGOR

Leena Sarah Farhat PRIFYSGOL BANGOR

Colin Jarvis OPENAI

Dewi Bryn Jones PRIFYSGOL BANGOR

Gruffudd Prys PRIFYSGOL BANGOR

Emily Barnes COLEG Y DRINDOD DULYN

Meinir Williams PRIFYSGOL BANGOR

Sarah Cooper PRIFYSGOL BANGOR

Stefano Ghazzali PRIFYSGOL BANGOR

Delyth Prys PRIFYSGOL BANGOR

Cynnwys

Rhagair

YR ATHRO EMERITA DELYTH PRYS, CYN-BENNAETH UNED TECHNOLEGAU IAITH CANOLFAN BEDWYR, PRIFYSGOL BANGOR

Pan ymunais i â Phrifysgol Bangor yn 1993 prin iawn oedd y ddealltwriaeth o'r rhan anferthol fyddai cyfrifiaduron, y we fyd-eang a chyfathrebu digidol yn ei gael ar fywydau pawb ymhen deg mlynedd ar hugain. Saesneg oedd iaith y cyfryngau hyn bron yn ddieithriad, ond yn araf fe ddaeth rhai o ieithoedd mawr eraill y byd yn fwy cyffredin arnynt. Roedd hi'n wahanol iawn i ieithoedd bach, llai eu hadnoddau, gan gynnwys ieithoedd mewn perygl a ieithoedd lleiafrifedig fel y Gymraeg. I'r ieithoedd hyn roedd dyfodiad y technolegau a'r cyfryngau digidol yn fygythiad, gan beryglu eu ffyniant a'u parhad. Roedd y technolegau newydd yn cyfleu pŵer a chyffro, cyfleoedd economaidd a dyfodol llewyrchus, gan wneud i'r ieithoedd oedd heb fynediad atynt edrych yn hen ffasiwn a llwm, ac yn llai tebygol o gael eu pasio ymlaen i'r genhedlaeth nesaf.

Dyma'r cefndir a barodd i griw bychan ohonom ym Mhrifysgol Bangor benderfynu mynd ati i ddatblygu technolegau iaith ar gyfer y Gymraeg, gan bartneru hyd yr oedd modd gyda diwydiant a'r sector gwirfoddol er mwyn i'r gymuned ehangach gael budd o'r ymchwil. Galluogi'r iaith Gymraeg i gael yr adnoddau angenrheidiol ar gyfer ffynnu yn y byd digidol oedd y nod, ac y mae hynny bellach yn dechrau cael ei wireddu mewn meysydd fel technoleg lleferydd a thechnoleg testun, a chyfieithu peirianyddol.

Ond y mae hwn yn fyd sy'n symud yn ei flaen yn gyflym, a hyd yn oed wrth gyhoeddi'r gyfrol gyntaf yng nghyfres Technoleg a'r Gymraeg yn 2021, prin yr oedd yr un ohonom yn meddwl am y chwyldro newydd a ddeuai yn fuan yn sgil deallusrwydd artiffisial cynhyrchiol fel ChatGPT. Gall y dechnoleg hon weithio ar draws ieithoedd dim ond cael digon o ddata testun a lleferydd i'w hyfforddi, ond mae 'digon' yma yn swm anferthol, yn cynnwys o leiaf biliwn o baramedrau yn ôl y syniadaeth gyfredol, rhywbeth sy'n anodd iawn i ieithoedd bach gyrraedd ato.

Yn ffodus, mae yna ffyrdd o weithio gyda'n gilydd ar draws cymunedau ieithyddol i ddysgu gan ein gilydd a chanfod technegau newydd i gyrraedd y nod. Mae hyn yn digwydd ar draws Ewrop ac ar draws y byd, ac yn nes adref, dechreuodd ymchwilwyr y Gymraeg a'r ieithoedd Celtaidd eraill gydweithio, gan ganfod nifer o resymau cymdeithasol, gwleidyddol a ieithyddol dros rannu profiadau a thechnegau.

Y mae ffrwyth peth o'r cydweithio hwnnw i'w weld yn y gyfrol yma, ac mae rhagor eto i ddod. Mae cefnogaeth gan yr Undeb Ewropeaidd, Llywodraeth Cymru a chyrff eraill wedi cynorthwyo technolegau iaith Celtaidd i aeddfedu, ac rwy'n edrych ymlaen i weld y datblygiadau nesaf yn y maes.

Diolch i'm holl gydweithwyr ym Mhrifysgol Bangor, ar draws Cymru ac yn ehangach, ac yn enwedig yn y byd technolegau iaith Celtaidd am bob cefnogaeth ac ysbrydoliaeth dros y deg mlynedd ar hugain diwethaf.

Rhagymadrodd

GARETH WATKINS – PRIFYSGOL BANGOR

Cyhoeddwn y gyfrol hon, yr ail yng nghyfres Iaith a Thechnoleg yng Nghymru, yn ystod cyfnod hynod gyffrous ym myd Technoleg Iaith. Nid yw Deallusrwydd Artiffisial (AI) yn gysyniad newydd,[1] fodd bynnag dros y flwyddyn ddiwethaf mae nid yn unig wedi tanio diddordeb academyddion, ond hefyd wedi dod yn rhan o fywydau bob dydd y cyhoedd. Yn ddiweddar, mae AI wedi bod yn bresennol yn aml yn y papurau newydd, yn bapurau poblogaidd a phapurau newydd safonol fel ei gilydd. Mae AI wedi'i wneud yn fwy hygyrch nag erioed o'r blaen. Mae'r sawl sydd â mynediad i'r rhyngrwyd yn gallu defnyddio AI trwy sgwrsfotiaid fel Copilot Microsoft a ChatGPT OpenAI. Mae cymwysiadau AI yn amrywiol ac yn niferus, heb fod wedi'u cyfyngu'n benodol i Dechnoleg Iaith nac i sgwrsfotiaid. Er enghraifft, mae AI yn cael ei ddefnyddio i wella cynhyrchiant cnydau,[2] i adnabod paentiadau ffug ar e-bay,[3] ac i uwchraddio a gwella delweddau teledu.[4]

Mae gan AI y potensial i wneud daioni mawr, ond os na fydd y dechnoleg hon ar gael i ieithoedd lleiafrifedig fe allai gael effaith ddinistriol o ran statws canfyddedig neu ddefnydd yr ieithoedd hynny. Mae angen i AI fod ar gael ar gyfer ieithoedd lleiafrifedig, felly, er mwyn sicrhau nad oes rhwystrau ieithyddol i ddefnyddio'r dechnoleg honno, ac fel nad yw siaradwyr ieithoedd lleiafrifedig yn troi cefn ar eu hiaith eu hun er mwyn gallu defnyddio AI.

Ond hyd yn oed heb ystyried yr effaith ar ddefnydd a statws iaith, mae angen datblygu'r dechnoleg hon ymhellach, a hynny mewn meysydd sy'n effeithio ar ieithoedd lleiafrifedig a rhai mwyafrifol fel ei gilydd. Gall ddioddef o ddrychiolaethau, lle mae'r system:

> "perceives patterns or objects that are non-existent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate." [1]

Fel yr adroddwyd yn eang, mae'r ynni a ddefnyddir, ac felly effaith amgylcheddol hyfforddi a defnyddio modelau AI, yn bryderus o uchel.[5] Mae gan hyd yn oed un defnydd unigol (neu *inference*) o'r dechnoleg gost ynni ac ôl troed carbon sy'n ofnadwy o uchel. Wrth drafod canlyniadau eu hymchwil, noda Luccioni et al. [2]:

> "the most efficient text generation model uses as much energy as 16% of a full smartphone charge for 1,000 inferences, whereas the least efficient image generation model uses as much energy as 950 smartphone charges (11.49 kWh), or nearly 1 charge per image generation."

Mae rhai yn pryderu am effaith AI ar swyddi ac ar yr economi, ac yn wir mae Georgieva [3] yn honni "AI will affect almost 40 percent of jobs around the world, replacing some and complementing others". Â ymlaen i honni y gallai'r ffigwr hwn gynyddu i 60 y cant mewn economïau datblygedig.

Yn amlwg, mae'n wir fod gwaith i'w wneud o hyd ar liniaru'r problemau hyn.

---

[1] Byddai'n werth i'r rhai sydd â diddordeb yn hanes hir AI ddarllen erthygl addysgiadol Rockwell Anyoha sydd ar gael o https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

[2] Gweler er enghraifft https://www.dmu.ac.uk/about-dmu/news/2024/april/dmu-using-ai-to-aid-crop-production-and-help-farmers-boost-income.aspx

[3] Fel yr adroddwyd gan y Guardian: https://www.theguardian.com/artanddesign/article/2024/may/08/fake-monet-and-renoir-on-ebay-among-counterfeits-identified-using-ai

[4] Gweler er enghraifft https://www.samsung.com/ca/support/tv-audio-video/how-to-use-the-intelligent-mode-of-samsung-qled-tvs/

[5] Gweler er enghraifft https://www.theguardian.com/technology/2024/mar/07/ai-climate-change-energy-disinformation-report

Mae hefyd yn wir bod y dechnoleg a ddaeth cyn ffyniant AI yn parhau i fod yn berthnasol. Mae ymchwil yn parhau ym meysydd Prosesu Iaith Naturiol (NLP) neu Ieithyddiaeth Corpws, er enghraifft. Mae hefyd yn bwysig cynnal allbynnau ymchwil blaenorol fel bod yr allbynnau hynny'n parhau'n berthnasol ac yn ddefnyddiol. Er gwaethaf y datblygiadau diweddar mewn AI, erys hen heriau, yn fwyaf nodedig, efallai, o ran argaeledd a chasglu data ieithoedd lleiafrifol.

Mae'r gyfrol hon yn clymu ynghyd sawl agwedd amrywiol ar Dechnoleg Iaith yng nghyd-destun yr iaith Gymraeg ac ieithoedd lleiafrifol eraill. Fel gyda'r gyfrol flaenorol, mae'r gyfrol hon yn gyfraniad i ddatblygiad y maes yng Nghymru, ac fe'i cyhoeddir yn ddwyieithog o dan drwydded agored (yn benodol CC-BY 4.0) fel y gall eraill ei defnyddio o dan delerau'r drwydded honno. Bydd hefyd yn cael ei hychwanegu at gorpws trwydded ganiataol Bangor, gan gyfrannu mwy o ddata y gellir ei ddefnyddio i ddatblygu Technoleg Iaith.

Mae llawer wedi newid ers cyhoeddi Iaith a Thechnoleg yng Nghymru Cyfrol 1. Mae llawer yn debygol o newid yn y blynyddoedd i ddod. Ac eto mae llawer wedi aros yr un peth, a bydd llawer yn parhau i aros yr un peth. Rhagwelir y bydd y gyfrol hon yn helpu i addysgu ac ysbrydoli ymchwilwyr y dyfodol a'r presennol, a thrwyddynt hwy, gyfrannu at ddatblygiad Technoleg Iaith yn y dyfodol.

CYFEIRIADAU

[1] IBM. 2024. What are AI hallucinations? Adalwyd o https://www.ibm.com/topics/ai-hallucinations

[2] Alexandra Sasha Luccioni, Yacine Jernite ac Emma Strubell. 2023. Power Hungry Processing: Watts Driving the Cost of AI Deployment? Tach. 28 2023. arXiv:2311.16863v1.

[3] Kristalina Georgieva. 2024. AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity. Adalwyd o https://www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity

Gwnewch Ramadegau Wici!

MÉLANIE JOUITTEAU

l'Université Bordeaux Montaigne a l'Université de Pau et des Pays de l'Adour

LOÏC GROBOL

Université Paris Nanterre

Yn y bennod hon ceir gwerthusiad o bersbectif prosesu iaith naturiol (NLP) o'r cysyniad o ramadegau wici, (wikigrammars), gan ddefnyddio gramadeg wici Llydaweg ARBRES fel astudiaeth achos. Mae'n archwilio'r defnydd o lwyfan wedi'i sylfaenu ar wici ar gyfer dogfennu amrywiaeth gystrawennol iaith Geltaidd prin ei hadnoddau, gyda chydran ryngweithiol wedi'i hanelu at gael y gymuned i gymryd rhan yn y gwaith. Mae'n cynnwys corpws cynhwysfawr wedi'i anodi sy'n bwydo i ieithyddiaeth theoretig ac i NLP. Dadleuwn yma dros fabwysiadu llwyfannau o'r fath gan gymunedau sy'n siarad ieithoedd lleiafrifedig, gan ddadlau eu bod yn darparu corpora gydag amrywiaeth gyfoethog o ran cystrawen, orgraff ac arddull. Gall amrywiaeth gramadegau wici wedi'u dethol yn artiffisial helpu ychydig gyda phrinder corpora helaeth sydd ar gael yn ddilyffethair mewn cyd-destunau ieithoedd prin eu hadnoddau.

Allweddeiriau: Ramadegau Wici, NLP, Llydaweg, Corpws

1 CYFLWYNIAD

Mae gramadeg wici yn llwyfan seiliedig ar wici sy'n disgrifio iaith ac sydd ar gael yn agored i gyfraniadau a thrafodaethau, lle mae'r enghreifftiau wedi'u hanodi ac mae modd eu cyrchu yn awtomatig.

Mae gramadegau wici yn darparu math penodol iawn o ddatblygiad technoleg iaith: corpws sydd drwy ddiffiniad yn gyddwysiad o amrywiaeth ieithyddol. Yn yr erthygl hon, rydym yn cyflwyno rhai nodweddion gramadeg wici ARBRES o dafodieithoedd Llydaweg [1], iaith Geltaidd brin ei hadnoddau. Rydym yn argymell i gymunedau o ieithoedd lleiafrifedig ddefnyddio'r datrysiad hwn er mwyn meithrin datblygiad eu hecosystem adnoddau digidol ar gyfer technolegau iaith.

2 AMRYWIAETH IEITHYDDOL DRWY DDYLUNIAD

Prif nod ARBRES yw darparu disgrifiad cynhwysfawr o'r Lydaweg, gan ddal ei hamrywiaeth, cymhlethdod a rheoleidd-dra, mewn ffordd hygyrch yn ei ffurfiau ar-lein ac ysgrifenedig ar gyfer y gymuned ieithyddol. Dylai gramadeg o'r fath nid yn unig enwi a disgrifio'r strwythurau mwyaf cyffredin, ond hefyd eithriadau a ffenomenau anfynych. Mae dosbarthiad ystadegol geiriau a strwythurau felly yn llawer mwy amrywiol nag a fyddent mewn hap sampl tua'r un faint o'r iaith. Mae ffactor arall yn cyfrannu at amrywiaeth y data: am resymau hawlfraint, ni allai'r awdur ond cymryd cyfran fach o'r brawddegau ar gyfer pob corpws cyhoeddedig. Effaith hyn yw lledu'r amrywiaeth ffynonellau a geir ar gyfer corpora rhad ac am ddim wedi'u hargraffu (llenyddiaeth, erthyglau papur newydd, nofelau, caneuon, cerddi, casgliadau o ymadroddion poblogaidd, taflenni gwleidyddol, gwefannau gwybodaeth neuaddau tref, postiadau ar rwydweithiau cymdeithasol, ac ati).

Ail nod ARBRES yw darparu adnodd wedi'i ddogfennu ar gyfer trafodaethau cyfredol mewn ieithyddiaeth theoretig. Yn y ffordd honno, mae'n debyg i nodiadur ymchwil arferol ar gyfer ei brif awdur. Mae hyn hefyd wedi cael effaith ar y data deilliannol: mae'n cynnwys y brawddegau artiffisial braidd sy'n nodweddu gramadegau a

phapurau ymchwil. Fodd bynnag, caiff hyn ei orbwyso gan lawer iawn mwy o enghreifftiau mwy naturiol. Mae'r ffynhonnell hon o ddata yn cynnwys parau lleiafsymiol a thystiolaeth negyddol.

Ffynhonnell bwysig arall o enghreifftiau yw data enyn gwybodaeth mewn gwaith maes, lle mae siaradwyr brodorol wedi cael protocolau cwestiynau cyfieithiadau neu tasgau disgrifiadol o ddelweddau. Caiff canlyniadau amrwd y gweithgaredd ennyn gwybodaeth eu cyhoeddi ar-lein ac maent yn bwydo'r gramadeg. Mae'r protocolau hyn yn cynnwys tasgau barnu pa mor ramadegol yw brawddegau, gan gynhyrchu enghraifft anramadegol sy'n gwasanaethau fel tystiolaeth negyddol gyferbyniol. Mae'r ffynhonnell hon o ddata hefyd yn cynnwys parau minimol a thystiolaeth negyddol.

2.1 Amrywiaeth dafodieithol a hanesyddol

Mae ARBRES yn ramadeg o dafodieithoedd, ac mae wedi'i gynllunio i gynnwys llawer o amrywiaeth dafodieithol. Mae'n ramadeg disgrifiadol, lle'r ystyrir Llydaweg safonol i fod ond yn un dafodiaith ymysg llawer. Mae'r sbectrwm tafodieithol felly yn eang iawn, gydag eithriad nodedig tafodiaith Bro Gwened, sydd yn ieithyddol y bellaf un oddi wrth y lleill, ac ar hyn o bryd heb gynrychiolaeth ddigonol yn ARBRES. Weithiau mae angen arbenigedd nad oes gan y prif olygydd er mwyn ei dadansoddi, ac o ganlyniad, mae llai o ddata ar gael ar hyn o bryd yn y dafodiaith hon.

Ar wahân i'r *caveat* hwn, gallwn ystyried, yn ystadegol, fod nodweddion tafodieithol prin yn cael eu cynrychioli yn ormodol yn y data. Yn wir, caiff nodweddion ieithyddol cyffredin eu darlunio gydag ychydig yn unig o enghreifftiau ar gyfer pob prif tafodiaith. I'r gwrthwyneb, er mwyn medru disgrifio pob nodwedd brin yn fanwl, gyda'i dosbarthiad tafodieithol a pharamedrau cyd-destun ei hymddangosiad, caiff pob enghraifft sydd yno eisoes ei hintegreiddio'n ofalus. Mae nodweddion prin hefyd yn fwy tebyg o fod yn destun ymchwil ennyn thematig, sy'n rhoi mwy o ddata lle mae'r nodweddion hyn yn digwydd. Er mwyn disgrifio'r amrywiad hefyd, bydd ffurfiau mewn arddulliau gwahanol yn cyd-fodoli o fewn y corpws, gyda'r amrywiad wedi'i or-gynrychioli yn feintiol o'i gymharu ag unrhyw gorpws unigol. Yn yr ystyr hwn, tra bod ARBRES yn llai dibynadwy o ran astudiaethau meintiol, mae'n addas iawn ar gyfer astudiaethau ansoddol.

Yn olaf, er nad yw ARBRES mewn gwirionedd yn waith diacronig, mae'n dal i gynnwys data o gyfnod Llydaweg Canol i Lydaweg yr unfed ganrif ar hugain. Mae presenoldeb data corpws ysgrifenedig o'r cyfnodau hyn yn awgrymu, yn enwedig ar gyfer yr ugeinfed ganrif, bresenoldeb nifer o systemau orgraffyddol sy'n cystadlu yn erbyn ei gilydd. Nid yw'r data ffynhonnell wedi cael ei newid, ac mae enghreifftiau yn ymddangos yn eu sillafiadau print gwreiddiol. Y canlyniad yw corpws gyda sawl orgraff wahanol ynddo.

2.2 Maint

Mae gwefan ARBRES wedi bod yn cael ei datblygu ers 2007, ac wedi cychwyn cael presenoldeb ar-lein yn 2009. Yn y 5 mlynedd diwethaf mae wedi cael mwy na 100 o ymwelwyr dynol y dydd. Erbyn mis Chwefror 2024, roedd 10,238 tudalen arni, yn cynnwys 4,804 tudalen o gynnwys, 19 tudalen o gyflwyniadau, a nifer o dudalennau ailgyfeirio. Ceir 3,094 erthygl ar elfennau o ramadeg Llydaweg a 325 dalen esboniadol ar gwestiynau yn ymwneud â theori yn y tudalennau cynnwys. Gyda'i gilydd, mae hyn yn gyfanswm o tua 15k brawddeg wreiddiol mewn Llydaweg, wedi'u glosio a'u cyfieithu i'r Ffrangeg, yn dod o 1,208 o weithiau ymchwil ar yr iaith Lydaweg (llyfrau, geiriaduron, erthyglau ymchwil, blogiau casgliadau ymchwil), 493 cyfeiriad corpws wedi'u cynhyrchu gan siaradwyr brodorol (gan mwyaf mewn ffurfiau ysgrifenedig: nofelau, erthyglau papur newydd, caneuon) a 44 sesiwn ennyn gyda siaradwyr brodorol.

3 FFYNHONNELL DATA AR GYFER NLP

3.1 Anodiadau morffogystrawennol

Rhoddir yr enghreifftiau ar ffurf tablau wici (wikitables), tablau yn iaith tagio meddalwedd MediaWiki [2] sy'n pweru ARBRES. Mae pob un o'r tablau hyn yn darparu ar gyfer alinio un frawddeg unigol o'r geirffurfiau gwreiddiol a'u glosiau, cyfieithiadau o'r frawddeg yn ei chyfanrwydd, enw'r dafodiaith, a chyfeiriad y ffynhonnell. Mae pob glos eirffurf yn cael ei gysylltu drwy hyperddolen at dudalen benodedig, gan gynnwys o leiaf ei lema safonol a'i gategori gramadegol. Oherwydd yr amlder sillafiadau, mae hyn yn galluogi cysondeb sylweddol yn y data heb amharu ar yr amrywiaeth.

Mae'r system hon hefyd yn ei gwneud hi'n bosib cyrraedd yr holl ffurfiau ar gyfer unrhyw lema, sy'n hanfodol i'r iaith Geltaidd hon, lle mae ffurfdroadau nid yn unig yn cynnwys newidiadau i ddiwedd geiriau ond hefyd addasiadau i'r gytsain gyntaf yn dibynnu ar y cynnwys cystrawennol (treigliadau cytseiniol). Gellir felly gysylltu'r lema *krokodil* yn awtomatig i'r enghreifftiau ohono yn *krokodil Maia* (crocodil Maia), *ar c'hrokodil* (y crocodil), *ar c'hrokodiled* (crocodilod) a *war grokodileta* (ar fin edrych am grocodilod), yr holl enghreifftiau hyn yn pwyntio at y dudalen am y lema *krokodil*. I'r gwrthwyneb mae tudalennau dadamwyso yn darparu rhestrau clicadwy o forffemau a geiriau gyda mwy nag un ystyr.

O safbwynt technoleg iaith, golyga hyn fod y glosau ar ARBRES eisoes yn gorpws wedi'i anodi yn forffogystrawennol; set o frawddegau, gyda lemata a thagiau rhannau ymadrodd ar gyfer pob gair a nodweddion morffolegol ychwanegol. Mae hefyd yn gwneud hedyn da iawn ar gyfer tyfu [3]. Am fanylion ychwanegol am yr anodiadau gramadegol adferadwy gweler Jouitteau a Bideault [4].

3.2 Data cyfochrog

Mae'r glosau i gyd yn cynnwys cyfieithiadau i'r Ffrangeg, yn dod naill ai o'u cyhoeddiad gwreiddiol neu wedi'u darparu gan yr awdur, ym mhob achos gan siaradwyr Llydaweg rhugl. Er bod y cyfieithiadau hyn wedi'u darparu yn wreiddiol i helpu pobl nad oedd yn medru'r Llydaweg i wneud synnwyr o'r deunydd ffynhonnell, gellir edrych arnynt hefyd fel corpws cyfochrog o frawddegau.

Mae'r corpws hwn yn eithaf bach o ran maint, ond mae o ansawdd da iawn ac mae iddo amrywiaeth llawer mwy nag a fyddai gan hap sampl o faint cymharol. Mae ei ansawdd yn deillio yn syml o darddiad y data: mae'r holl frawddegau wedi cael eu dewis â llaw, eu cyfieithu gan siaradwyr rhugl, a'u dilysu yn ofalus i sicrhau eu perthnasedd i ddarlunio ffenomenau ieithyddol. Sicrheir yr amrywiaeth eang gan swyddogaeth y glosau, gan eu bod wedi'u dethol i enghreifftio gymaint o ffenomenau ieithyddol â phosibl, bydd ffenomenau prin yn fwy niferus ynddo o'i gymharu â faint fyddai'n digwydd yn naturiol.

Mae arbrofion sy'n digwydd ar hyn o bryd i ddatblygu system cyfieithu peirianyddol gan ddefnyddio allforiad cynnar o'r data hwn (tua 5,000 o frawddegau ar ôl dileu dyblygion, data negyddol a data lle methodd yr allforiad) yn tueddu i gadarnhau bod y nodweddion hyn yn gwneud ARBRES yn set ddata werthfawr iawn. Yn wir, mae ei gynnwys yn y data hyfforddi ar gyfer systemau oddi-ar-y-silff yn rhoi cynnydd mewn perfformiad sy'n debyg i'r hyn a geir gyda llawer iawn mwy o ddata [5].

3.3 Amcangyfrifon cost

Mae defnyddio gramadegau wici fel ffynonellau data ieithyddol yn ddrud, gan ei fod yn golygu un neu fwy o bobl sydd wedi'u hyfforddi yn yr iaith, gyda rhywfaint o hyblygrwydd tafodieithol, a llwyfan cymdeithasol sy'n addas i

gyrraedd siaradwyr o broffiliau ieithyddol gwahanol. Mae hefyd angen cefnogaeth dechnegol i ddylunio a chynnal y wefan, a sicrhau ei bod yn hygyrch. Mae angen gweithwyr cymwys hefyd i alldynnu'r data. Y dasg fwy llafurus yw codio eang ar yr enghreifftiau i'w cyflwyno yn addas o fewn y tablau wici. Mae cymhlethdod y dasg hon yn esblygu'n sydyn, oherwydd gwelliannau mewn cynhyrchu iaith naturiol. Ar gyfer y gramadeg wici, cymerodd 15 mlynedd i un anodwr ar ei phen ei hun, yn gweithio tua hanner amser arno, i brin gyrraedd anodi 15,000 o frawddegau.

Mae sgwrsfotiaid yn awr yn galluogi awtomeiddio rhan sylweddol o'r gwaith anodi. Er enghraifft, ers 2024, gyda phromptiau digon manwl yn rhoi saith enghraifft o ddata strwythuredig, gall Chat GPT 3.5 ddosbarthu tocynnau ar draws tablau, alinio glosau, amgodio rhan fawr o ddolenni cliciadwy, cynnig cyfieithiadau (heb fod yn gywir bob tro, ond wedi'u halinio'n gywir), a threfnu cyfeiriadau ffynhonnell yn gywir. Mae'n dal yn hanfodol cael arbenigwr dynol yn rhan o'r broses, ond mae wedi cael ei symleiddio'n rhyfeddol, i'r fan lle gall unigolyn yn hawdd fewnbynnu 300 enghraifft y dydd. Mae ChatGPT 4 yn gwella'r broses hon ymhellach gyda chyfieithiadau o well safon. Wrth gwrs, mae'r gallu olaf hwn yn dibynnu ar faint ac ansawdd data'r iaith darged o fewn set data hyfforddi ChatGPT. Mae gan y systemau hyn ddiffygion hysbys, yn arbennig o ran effaith cymdeithasol ac aneffeithiolrwydd (gweler Solaiman et al. [6] a chyfeiriadau o'i fewn), ond mae eu gwerth fel offer cynorthwyol ar gyfer y dasg hon yn dangos yn eithaf da faint y gallai systemau wedi'u datblygu'n unswydd ar gyfer y dasg hon eu cyflawni (tra'n osgoi'r diffygion a enwyd).

Yr hyn sy'n newydd yn y datrysiad hwn yw y gallai'r holl adnoddau ac amcanion hyn fodoli y tu allan i sgôp ymchwil NLP. Gall y buddsoddiad gael ei yrru yn gyfan-gwbl gan amcanion mewnol ar lefel y gymuned, neu gan ddibenion ieithyddol neu wyddonol. Ar ben hyn, ysgrifennwyd ARBRES gan ieithydd ffurfiol, ond does dim rhaid i hynny fod; cyn belled a bod y gramadeg wedi'i ysgrifennu i addysgu bodau dynol am yr iaith, bydd y swm angenrheidiol o amrywiaeth ar gael yn y data. Gellir wedyn adeiladu'r adnodd ychydig bach ar y tro fel adnodd addysgol a/neu wyddonol mewn ffurf sydd wedi'i haddasu ar gyfer ei gynulleidfa. Ar raddfa cymunedau ieithyddol bach, mae hyn yn osgoi monopoleiddio arbenigwyr i greu adnoddau na fyddai modd eu defnyddio gan y cyhoedd yn gyffredinol. Mae anodiadau mwy arbenigol y data (categoreiddio gramadegol, lemateiddio, codio treigliadau cytseiniaid) yn parhau i fod o'r golwg ynddynt, a dim ond yn gymorth llywio i'r darllenydd dynol.

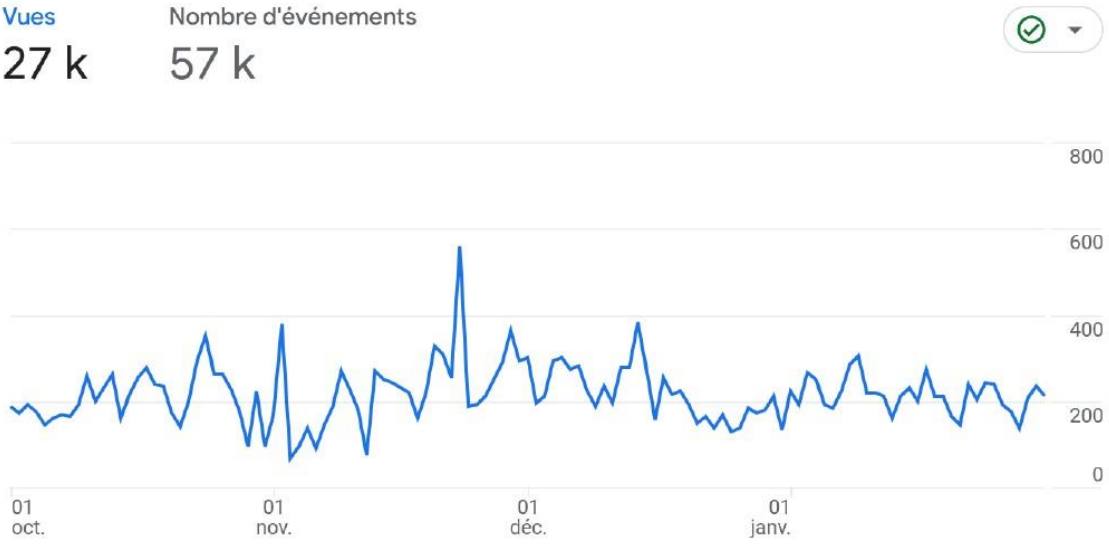
Argymhellir datblygu gramadegau wici yn arbennig ar gyfer adeiladu adnoddau projectau peilot i ieithoedd gyda chorpora cyfyngedig, oherwydd hyd yn oed lle mae gweithredwyr technoleg am y tro yn methu darparu offer gorffenedig ar gyfer siaradwyr, bydd y buddsoddiad yn parhau yn werthfawr ar gyfer y gymuned o siaradwyr, a fydd mewn gwirionedd yn medru parhau i wella'r gramadeg wici ar ei chyfer ei hun.

Mae ieithyddion disgrifiadol a ffurfiol yn gosod i'w hunain y dasg o gynhyrchu deunydd dadansoddi iaith, a gallant ddatblygu'r rhain heb hyfforddiant NLP. Mae gan gystrawen y wici gost mynediad isel iawn, sydd bellach tua'r un gost â rhaglen prosesu geiriau arferol. Mewn ieithoedd gyda chorpora cyfyngedig, mae ieithyddion ac arbenigwyr hyfforddedig yn aml yn ymrwymedig iawn i'w parth empirig ac i'r siaradwyr sy'n cynhyrchu'r data. Fel arfer mae ganddynt wybodaeth ddiwylliannol fanwl, yn cynnwys amrywiaeth data byw, ac mae gan hyn hefyd effaith gadarnhaol ar yr enghreifftiau a ddewiswyd. O ran adnoddau dynol, mae'r datrysiad hwn yn ei gwneud hi'n bosibl dal gafael yn eu harbenigedd manwl. Mae hyn yn arbennig o addas ar gyfer ieithoedd lleiafrifedig lle mae ieithyddion ac arbenigwyr hyfforddedig fel arfer yn brin o ran nifer, ac weithiau mewn sefyllfaoedd socioeconomaidd bregus. Yn olaf, mae gramadegau wici yn galluogi'r corpws i gael ei adeiladu o dan adolygiad, uniongyrchol ac anuniongyrchol, y gymuned gyfan o siaradwyr.

4 YMGYSYLLTU CYMDEITHASOL MEWN IEITHOEDD LLEIFAFRIFOL

4.1 Ymgysylltiad cyhoeddus

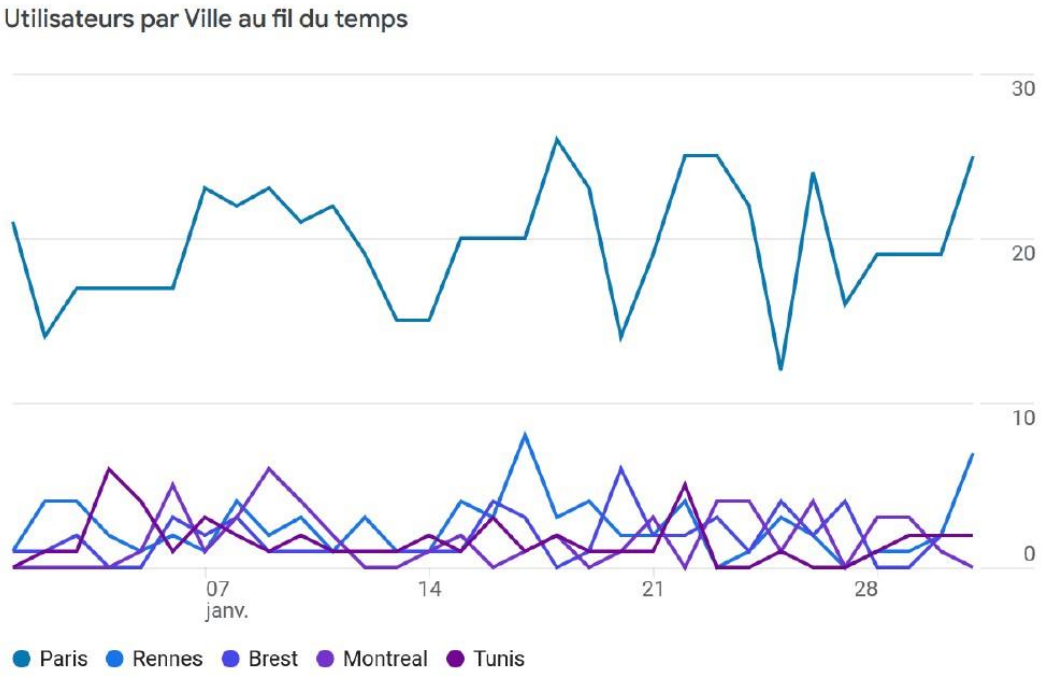
Mae offer ystadegol mewnol, yn ogystal â systemau dadansoddi allanol yn rhoi gwybodaeth fanwl am ddefnydd y wefan, gan dracio (heb fanylion defnyddwyr) y 100 a mwy o ymweliadau dynol dyddiol ag ARBRES. Mae'r graff yn Ffigur 1 yn dangos ystadegau ymweliadau o bob man drwy'r byd o Hydref 2023 hyd at ddiwedd Ionawr 2024.



Ffigur 1: Nifer yr ymweliadau ar ARBRES o Hydref 2023 I ddiwedd Ionawr 2024.

Mae astudio llif y darllenwyr yn ei gwneud hi'n bosib adnabod a deall bylchau. Gall rhywun weld y tudalennau cofnodi llwyddiannus, y rhai sy'n cael llai o ymgysylltiad neu'r amser darllen byrraf, neu'r ceisiadau penodol wnaed ar beiriannau chwilio sy'n arwain darllenwyr at y gramadeg.

Unwaith i'r wefan gyrraedd maint critigol a bod cynrychiolaeth dda ohoni mewn peiriannau chwilio, mae modd dadansoddi ffynonellau daearyddol cysylltiadau i ddarparu gwybodaeth ar y darllenwyr. Defnyddir ARBRES yn bennaf o fewn Llydaw ac o fewn cymunedau ar wasgar, fel y gwelir yn Ffigur 2, sy'n adrodd y nifer o ymweliadau fesul dinas yn Ionawr 2024.



Ffigur 2: Nifer yr ymweliadau ar ARBRES yn Ionawr 2024 fesul dinas

Mae defnydd y wefan yn nodedig o agos at y calendr academaidd. Mae'r adrannau sy'n delio ag agweddau mwy cymhleth ieithyddiaeth ffurfiol, sy'n cynnig gwybodaeth sylfaenol mewn Ffrangeg, yn profi cynnydd mawr yn y traffig yn ystod cyfnodau arholiadau arferol rhanbarthau Ffrangeg eu hiaith (e.e. y Swistir, Moroco, Québec, Algeria, Gwlad Belg, ac ati).

Mae manyldeb y data daearyddol yn galluogi edrych ar ddefnydd rhyngwladol yr adnodd, megis pan fydd cyrsiau Llydaweg yn cyfeirio ato. Er enghraifft, yn 2010, dechreuodd Anna Mouradova ddysgu Llydaweg yn Moscow, a esgorodd hyn ar gynnydd sydyn yn y cysylltiadau. Yn ddiddorol, mae modd gweld hefyd lle nad yw'r adnodd yn cael ei adnabod (megis y dosbarthiadau Llydaweg ysbeidiol yn Harvard).

4.2 Arfogi'r rhyngwyneb rhwng gwyddoniaeth a chymdeithas

Mae'r gramadeg wici Llydaweg ARBRES yn arbrawf mewn gwyddoniaeth agored a chyfranogol (gweler Jouitteau [7] am ddadansoddiad cynnar o'r cynnyrch). Mae gramadegau wici yn dod â'r broses wyddonol yn nes at y cyhoedd. Fel unrhyw ramadeg arall sydd â mynediad agored, mae'n darparu canlyniadau'r ymchwil ar ddiwedd ei broses ar amser penodol. Ond mae'n gwneud llawer iawn mwy na hynny. Ar yr un pryd, mae'n cydio'r gwaith wrth y ffynonellau a ddefnyddiwyd ac wrth y gymuned wyddonol. Mae hefyd yn taflu goleuni ar orffennol ei wneuthuriad, ac ar ddyfodol ei wneuthuriad. Fe wnawn ni yn awr ddarlunio'r tri dimensiwn hwn.

Mae monitro gwyddonol yn ei gwneud hi'n bosib bwydo'r gramadeg gyda chanlyniadau'r ymchwil diweddaraf. Mae'r effaith hon yn dod yn unig o'i ddefnydd fel nodiadur ymchwil. Caiff yr adnoddau allanol eu crynhoi, eu cyfeirio

atynt, a lle mae mynediad agored yn caniatáu hynny, mae'n rhoi dolen uniongyrchol atynt. Mae'r holl weithrediadau hyn yn dod â'r darllenwyr yn nes at y rhanddeiliaid gwyddonol, gan eu gwneud yn fwy dealladwy ac yn fwy hygyrch. Yn 2014, gofynnodd trefnwyr y Redadeg (digwyddiad Ras ar gyfer y Lydaweg) am gyfieithiad o "Rwy'n siarad Llydaweg, beth amdanat ti?' mewn gwahanol ieithoedd. Ymhen ychydig ddyddiau, cyfrannodd ieithyddion o bob rhan o'r byd yn barod iawn i'r dudalen Rwy'n siarad Llydaweg, beth amdanat ti?, gan gyfrannu cyfieithiadau o'r frawddeg hon i 77 iaith wahanol. I gefnogi'r achlysur, postiodd 1,695 siaradwr Llydaweg eu hunan-bortread ar-lein gyda'r brawddegau hyn. Gwnaed y gymuned ryngwladol o ieithyddion yn weladwy i'r gymuned, ac o'r tu arall gwnaed yr iaith Lydaweg mewn modd diriaethol iawn yn gynhyrchiad siaradwyr byw i'r gwyddonwyr.

Mae gramadeg wici hefyd yn cynnwys ei holl hanes. Mae'n cyfeirio at wneud ei ymchwil ei hun. Mae'r ffwythiant hanes wici yn caniatáu olrhain y broses o adeiladu gwybodaeth a chasglu data yn llawn: cyfraniadau, cywiriadau, trafodaethau, archwilio setiau data newydd, integreiddiad ffynonellau llyfryddol newydd a rhagdybiaethau newydd sy'n codi ac yn cael eu profi. Mae pob tudalen yn cael ei chysylltu gyda hanes llawn sy'n rhoi'r holl newidiadau a wnaed ers ei chreu. Gall rhywun olrhain sut mae gwyddoniaeth yn cael ei gyflawni, sut mae data newydd a chyhoeddiadau newydd yn newid ein rhagdybiaethau. Mae amrywiaeth cyfranwyr neu ddiffyg hynny ar gyfer pob pwnc yn weladwy. Mae pob cyfraniad yn weladwy ac mae modd rhoi cydnabyddiaeth i bob un.

Mae ymchwil gwyddonol yn ganlyniad methodoleg, ac yn y bôn mae hynny'n broses sy'n hygyrch i bawb, cyhyd â bo'r fethodoleg yn cael ei pharchu. O fewn y cyfyngiadau hyn mae'r feddalwedd wici wedi'i chynllunio i ganiatáu cydweithio cynyddol (agregu llawer iawn o gyfraniadau bach i un bensaernïaeth), a chydweithrediad dosbarthol (gyda thasgau gwahaniaethol). Gall gwahanol gymwyseddau wedyn ddod at ei gilydd i adeiladu adnodd cryf ar gyfer y gymuned. I'r darllenydd mae'r cyfrwng hwn yn codi'r cwestiwn o'i le yn y broses, gan alluogi graddau gwahanol o swyddogaethau o'r goddefol i'r gweithredol (darllen, rhoi sylwadau, cywiro, darparu mewnbwn, ysgrifennu, cydgysylltu, ac ati). Mae croeso arbennig i hyn yn achos ieithoedd lleiafrifol, lle mae siaradwyr yn gyffredin yn teimlo bod eu hiaith yn cael ei dwyn oddi arnynt.

Yn olaf, gadewch i ni drafod effaith ymylol ond llesol. Mae cymdeithas yn frith o drafodaethau safon gwael am ieithoedd ac yn enwedig ieithoedd lleiafrifol, oherwydd diffyg gwybodaeth y mae modd ei gwirio, prinder gwybodaeth wrthrychol am amrywiadau iaith, neu groniadau o ddiffyg cywirdeb. Mae'r gramadeg wici yn lletya erthyglau yn trafod iaith sy'n cyflwyno elfennau diriaethol o ddadansoddi i'r trafodaethau hyn, a chyfeiriadau gwyddonol go iawn. Mae fformat digidol yr erthyglau hyn yn peri bod modd eu rhannu yn uniongyrchol ar rwydweithiau cymdeithasol, mewn fformat sy'n agored i drafodaeth wyddonol, o fewn cyfyngiadau dadleuon gwyddonol. Yn ARBRES, yr erthygl am ragdybiaeth Sapir-Whorf yw'r ail dudalen fwyaf poblogaidd o ran nifer yr ymweliadau ar y wefan.

Yn ei dro, mae'r traffig y mae hyn yn ei gynhyrchu yn cynnal optimeiddiad peiriannau chwilio ac yn cynorthwyo gweledded gwaith sy'n ymwneud â ieithoedd wedi'u hymyleiddio ar y rhyngrwyd.

CYFEIRIADAU

[1] Mélanie Jouitteau. 2009–2024. ARBRES, Wikigrammaire Des Dialectes Du Breton et Centre de Ressources Pour Son Étude Linguistique Formelle. Adalwyd o http://arbres.iker.cnrs.fr

[2] Magnus Manske a Lee Daniel Crocker. 2002. MediaWiki. Adalwyd o https ://www.mediawiki.org/wiki/MediaWiki

[3] Mélanie Jouitteau, Yidi Jiang, Yingzi Liu, Salomé Chandora, Kim Gerdes, Bruno Guillaume, Adrien Said-Housseini a Sylvain Kahane. 2022–2024. Autogramm/Breton II. Adalwyd o https://github.com/Autogramm/Breton

[4] Mélanie Jouitteau a Reun Bideault. 2023. Outils Numériques et Traitement Automatique Du Breton. Yn: Langues Régionales de France: Nouvelles Approches, Nouvelles Méthodologies, Revitalisation. Société Linguistique de Paris, 37–74.

[5] Loïc Grobol, a Mélanie Jouitteau. 2024. ARBRES Kenstur: A Breton-French Parallel Corpus Rooted in Field Linguistics. Yn: I ddod.

[6] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker,Su Lin Blodgett, Hal Daumé III au2, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait a Apostol Vassilev. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. Meh. 12, 2023. arXiv: 2306.05949.

[7] Mélanie Jouitteau. 2012. La linguistique comme science ouverte. Yn: Lapurdum. Euskal ikerketen aldizkaria | Revue d'études basques | Revista de estudios vascos | Basque studies review 16 (16 1af Hyd. 2012), 93–115. doi: 10.4000/lapurdum . 2357.

Crynhoi Testun Awtomatig ar gyfer y Gymraeg

JONATHAN MORRIS

Prifysgol Caerdydd

IGNATIUS EZEANI

Prifysgol Caerhirfryn

IANTO GRUFFYDD

Prifysgol Bangor

KATHARINE YOUNG

Prifysgol Caerdydd

LYNNE DAVIES

Prifysgol Caerdydd

MAHMOUD EL-HAJ

Prifysgol Caerhirfryn

DAWN KNIGHT

Prifysgol Caerdydd

Mae crynhoi testun yn ddull digidol o grynhoi gwybodaeth 'allweddol' a geir o fewn testunau, a chreu fersiynau wedi'u talfyrru o destunau yn seiliedig ar y cynnwys hwn. Swyddogaeth crynhoi data yw darparu crynodebau ystyrlon, byr i ddefnyddwyr, rhywbeth sy'n aml yn cymryd amser ac yn anodd ei wneud â llaw. Mae hyn yn ddefnyddiol yn y byd modern digidol lle mae creu a rhannu testun yn cynyddu drwy'r amser, gan ei fod yn galluogi defnyddwyr i ganfod eu ffordd yn hawdd drwy'r llwyth o wybodaeth ddigidol sydd ar gael, a gwneud synnwyr ohono. Mae'r bennod hon yn adrodd am waith ar broject sy'n ceisio datblygu offeryn Crynhoi Testun Awtomatig ar gyfer y Gymraeg, ACC (Adnodd Creu Crynodebau). Mae'r bennod hon yn rhoi cyd-destun i'r angen am yr offeryn crynhoi testun hwn, yn egluro sut y crëwyd set ddata ar gyfer hyfforddi a phrofi'r dulliau, ac yn amlinellu cynlluniau ar gyfer datblygu'r crynhöwr.

Allweddeiriau: Crynhoi Testun, Cymraeg, alldynnu, creu a gwerthuso set ddata

1 RHAGARWEINIAD

Mae hanes hir i'r gwaith ar grynhoi testun awtomatig mewn Prosesu Iaith Naturiol. Ar y dechrau, roedd y gwaith hwn yn canolbwyntio yn unig ar Saesneg, fel lingua franca fyd-eang, ond yn awr caiff ei defnyddio mewn amryw o gyd-destunau gyda ieithoedd eraill, gan gynnwys Ffrangeg, Sbaeneg, Hindi, ac Arabeg, ymhlith eraill. Mae'r project 'MultiLinga' a chyfres gynadleddau sy'n cyd-fynd ag ef, er enghraifft, yn nodedig am gefnogi datblygu crynhoi testun mewn nifer o'r 7000+ gwahanol ieithoedd geir yn y byd. Mae'r wefan, http://multiling.iit.demokritos.gr yn darparu storfa agored ar gyfer data profi/hyfforddi tasgau crynhoi a chrynodebau model, ymhlith eraill.

Yr hyn sydd ar goll o'r adnoddau crynhoi presennol yw offer sy'n gweithio'n effeithiol gyda'r Gymraeg. Mae datblygiad ACC yn cyfrannu gwaith ar grynhoi testun mewn ieithoedd lleiafrifol ac yn ychwanegu at yr adnoddau technolegol sydd ar gael i siaradwyr Cymraeg.

2 Y GYMRAEG AR-LEIN

Cymharol isel yw'r defnydd o wefannau ac e-wasanaethau Cymraeg, ar waethaf y ffaith fod nifer o arolygon yn awgrymu yr hoffai siaradwyr Cymraeg gael mwy o gyfleoedd i ddefnyddio'r iaith, a chafwyd hanes helaeth o anufudd-dod sifil er mwyn ennill hawliau ieithyddol yn y cyd-destun Cymraeg [1].

Un rheswm am y defnydd cymharol isel o opsiynau Cymraeg ar wefannau yw'r syniad y bydd yr iaith yn rhy gymhleth [1]. Nid peth newydd yw'r pryder am gymhlethdod gwasanaethau a dogfennau Cymraeg sy'n wynebu'r cyhoedd. Amlinellwyd cyfres o ganllawiau ar greu dogfennau hawdd eu deall yn Gymraeg yn Cymraeg Clir [2]. Noda Williams [2] fod yr angen am fersiynau wedi'u symleiddio yn fwy yn Gymraeg o bosibl nag ar gyfer Saesneg ag ystyried bod:

1. llawer o ddogfennau Cymraeg sy'n wynebu'r cyhoedd wedi'u cyfieithu o'r Saesneg
2. yr amrywiadau safonol o'r Gymraeg yn bellach i ffwrdd o dafodieithoedd lleol o'u cymharu â Saesneg
3. termau technegol newydd eu cyfieithu yn fwy tebygol o fod yn gyfarwydd i'r darllenwr.

Mae'r egwyddorion sy'n cael eu hamlinellu yn Cymraeg Clir felly yn cynnwys y defnydd o frawddegau byrrach, geiriau bob dydd yn hytrach na therminoleg arbenigol, a chywair iaith niwtral (yn hytrach na ffurfiol) [2].

Bydd ACC yn darparu'r modd i grynhoi a symleiddio ffynonellau iaith digidol fydd yn cynorthwyo i ateb ofnau siaradwyr Cymraeg fod yr iaith ar-lein yn rhy gymhleth.

Bydd ACC hefyd yn cyfrannu at isadeiledd digidol y Gymraeg. Un o brif themâu strategaeth ddiweddaraf Llywodraeth Cymru yw adfywiad y Gymraeg (ac yn arbennig isadeiledd digidol), ynghyd â chynyddu nifer y siaradwyr a'r defnydd o'r iaith [3]. Y nod yw "sicrhau bod y Gymraeg wrth wraidd arloesi mewn technoleg ddigidol i'w gwneud hi'n bosibl defnyddio'r Gymraeg ym mhob cyd-destun digidol" [3]. Yn dilyn cyflwyno Safonau'r Gymraeg (gw. [4]) ac ymdrech arbennig i fuddsoddi mewn technolegau Cymraeg a gwella'r ffordd mae dewis iaith yn cael ei gyflwyno i'r cyhoedd, bydd datblygu ACC yn cyd-fynd gyda'r gyfres o dechnolegau Cymraeg (e.e. [5]) ar gyfer crewyr cynnwys a darllenwyr y Gymraeg.

Rhagwelir hefyd y bydd ACC yn cyfrannu at addysg cyfrwng Cymraeg drwy alluogi addysgwyr i greu crynodebau i'w defnyddio yn yr ystafell ddosbarth fel offer pedagogaidd. Bydd crynodebau hefyd o ddefnydd i rai sy'n dysgu Cymraeg a fydd yn medru canolbwyntio ar ddeall y wybodaeth graidd o fewn testun.

3 ALLDYNNU DATA

Y cam cyntaf yn y broses ddatblygu yw datblygu corpws bach (set ddata) o ddata'r iaith darged a fydd wedyn yn cael ei grynhoi a'i werthuso gan anodwyr dynol a'i ddefnyddio i ddatblygu a hyfforddi'r modelau crynhoi awtomatig (h.y. gweithredu fel set ddata 'safon aur').

Dewiswyd y Wikipedia Cymraeg (Wicipedia)[1] fel y brif ffynhonnell o ddata ar gyfer creu set ddata Gymraeg ar gyfer ACC. Roedd hyn oherwydd bod nifer helaeth o destunau Cymraeg yn bodoli ar y wefan hon (dros 133,000 o erthyglau), i gyd ar gael o dan drwydded Dogfennu Rhydd GNU. Er mwyn sicrhau bod yr erthyglau yn cynnwys maint digonol o destun i'w alldynnu a'i ddefnyddio, sefydlwyd lleiafswm trothwy o 500 tocyn i bob erthygl a tharged yn cynnwys o leiaf 500 erthygl ar y dechrau. Ar y dechrau alldynnwyd detholiad o 800 tudalen Wicipedia a gyrchwyd fynychaf yn Gymraeg i'w defnyddio. Cynhwyswyd 100 tudalen Wicipedia ychwanegol o broject WiciAddysg[2] a drefnwyd gan Lyfrgell Genedlaethol Cymru a Menter Iaith Môn. Fodd bynnag, sylwyd nad oedd mwy

[1] Wikipedia Cymraeg: https://cy.wikipedia.org/wiki/Hafan (Wicipedia)
[2] WiciAddysg: https://cy.wikipedia.org/wiki/Categori:Prosiect_WiciAddysg

na 50% o'r rhestr wreiddiol hon o 100 tudalen Wicipedia yn cynnwys y trothwy o leiafswm o 500 tocyn. Er mwyn lleihau effaith hyn, defnyddiwyd rhestr o 20 o allweddeiriau Cymraeg i gynhyrchu 100 tudalen Wicipedia ychwanegol ar gyfer pob allweddair (wedi'u darparu gan yr awdur cyntaf ac a oedd yn cynnwys geiriau oedd yn gyfystyr i'r Gymraeg, hanes a daearyddiaeth Cymru). Ychwanegwyd hyn at y rhestr o 100 tudalen Gymraeg Wicipedia a olygwyd fwyaf a'r tudalennau o broject WiciAddysg.

Defnyddiodd y drefn alldynnu data broses ailadroddus syml a gweithredwyd sgript Python yn seiliedig ar yr APIWikipedia[3] sy'n cymryd tudalen Wicipedia; yn alldynnu'r cynnwys allweddol (testun yr erthygl, crynodeb, categori) ac yn gwirio p'un a yw'r erthygl yn cynnwys lleiafswm nifer o docynnau. Ar ddiwedd y broses hon, crëwyd y set ddata o gyfanswm o 513 tudalen Wicipedia oedd yn cwrdd â'r meini prawf a osodwyd. Mae'r set ddata a alldynnwyd yn cynnwys ffeil ar gyfer pob tudalen Wicipedia gyda'r strwythur a'r tagiau dilynol:

```
<title>Article Title on Wikipedia</title>

<text>Article Text</text>

<category>Article Categories</category>.
```

Mae'r ffeiliau hyn ar gael mewn fformatiau ffeil testun plaen a html.

4 CREU SET DDATA

Defnyddiwyd cyfanswm o 19 myfyriwr israddedig ac ôl-raddedig o Brifysgol Caerdydd i greu, crynhoi a gwerthuso'r set ddata. O'r myfyrwyr hyn, roedd 13 yn astudio ar gyfer gradd israddedig neu ôl-radd yn y Gymraeg oedd yn cynnwys hyfforddiant blaenorol ar greu crynodebau o destunau cymhleth. Roedd y chwe myfyriwr arall yn fyfyrwyr israddar raglenni gradd eraill yn y Dyniaethau a Gwyddorau Cymdeithas ym Mhrifysgol Caerdydd ac wedi cwblhau eu haddysg orfodol mewn ysgolion cyfrwng Cymraeg neu ddwyieithog.

Gofynnwyd i fyfyrwyr gwblhau holiadur cyn dechrau ar y gwaith er mwyn cael gwybodaeth fywgraffyddol amdanynt. Roedd cyfanswm o 17 myfyriwr wedi caffael y Gymraeg yn eu cartref. Roedd un myfyriwr wedi caffael yr iaith drwy addysg drochi cyfrwng Cymraeg ac roedd un myfyriwr wedi dysgu'r iaith fel oedolyn. Roedd y mwyafrif o fyfyrwyr yn dod o dde orllewin Cymru (n=11). Roedd yr ardal hon yn cynnwys siroedd Caerfyrddin, Ceredigion, Castell Nedd Port Talbot, ac Abertawe. Roedd pump myfyriwr arall yn dod o ogledd orllewin Cymru oedd yn cynnwys siroedd Môn a Gwynedd. Roedd un myfyriwr yn dod o dde ddwyrain Cymru (Caerdydd), un o ganolbarth Cymru (Powys), ac un o ogledd ddwyrain Cymru (Conwy).

Gellir gwahaniaethu'n fras rhwng Cymraeg y de a Chymraeg y gogledd. Mae'r ddau amrywiad iaith (y mae gwahaniaethau tafodieithol pellach yn bodoli o'u mewn) yn arddangos rhai gwahaniaethau ar bob lefel o strwythur iaith er bod modd i bob amrywiad ddeall ei gilydd. Gofynnwyd pedwar cwestiwn i fyfyrwyr wnaeth ennyn gwybodaeth am yr amrywiadau lecsigol, gramadegol a ffonolegol y byddent yn eu defnyddio fel arfer. Roedd y canlyniadau yn cyfateb yn fras i'r ardal ddaearyddol: roedd 11 myfyriwr yn defnyddio ffurfiau deheuol a saith myfyriwr yn defnyddio ffurfiau gogleddol (yn cynnwys y myfyriwr o ganolbarth Cymru). Roedd un myfyriwr, o Gaerdydd, yn defnyddio cymysgedd o ffurfiau gogleddol a deheuol.

Rhoddwyd cyfarwyddiadau llafar ac ysgrifenedig i fyfyrwyr ar sut i gwblhau'r dasg. Yn benodol, dywedwyd wrthynt mai nod y dasg oedd cynhyrchu crynodeb syml o bob erthygl Wicipedia (a ddyrannwyd iddynt) fyddai'n cynnwys y wybodaeth bwysicaf. Gofynnwyd iddynt hefyd gydymffurfio gyda'r egwyddorion canlynol:

[3] API Wikipedia: https://pypi.org/project/wikipedia/

- hyd pob crynodeb fyddai 230 - 250 gair
- dylid ysgrifennu'r crynodeb yng ngeiriau'r awdur ei hun a nid ei alldynnu (copïo a gludo) o'r erthygl Wicipedia
- ni ddylid cynnwys unrhyw wybodaeth nad oedd yn yr erthygl yn y crynodeb
- dylid anonymeiddio unrhyw gyfeiriad at berson byw yn yr erthygl yn y crynodeb (i gydymffurfio gyda gofynion moesegol pob sefydliad partner)
- dylid prawfddarllen pob crynodeb a'i wirio gan ddefnyddio meddalwedd gwirio sillafu (Cysill[4]) cyn cyflwyno

Rhoddwyd cyfarwyddyd pellach am y cywair iaith i'w ddefnyddio i greu'r crynodebau. Gofynnwyd i fyfyrwyr gydymffurfio'n fras â chanllawiau Cymraeg Clir [2] ac, yn arbennig i osgoi ffurfiau cryno'r ferf sy'n llai cyffredin, a'r modd goddefol, a defnyddio geirfa syml lle bo modd yn hytrach na thermau arbenigol.

Cwblhaodd pob myfyriwr rhwng 60 - 100 crynodeb rhwng Gorffennaf a Hydref 2021. Treuliwyd tua 30 munud ar gyfartaledd ar bob crynodeb. Mae'r set ddata gyflawn yn cynnwys 1,461 crynodeb gyda'r 39 crynodeb sy'n weddill heb eu cwblhau oherwydd i un myfyriwr dynnu allan o'r project yn gynnar a rhai enghreifftiau o erthyglau anaddas (e.e. rhestrau o bwyntiau bwled).

Gofynnwyd i dri o'r myfyrwyr ôl-radd hefyd werthuso'r crynodebau drwy roi sgôr o rhwng un a phump iddynt. Mae Tabl 1 yn dangos y meini prawf marcio.

Tabl 1: Meini prawf ar gyfer marcio'r crynodebau

Sgôr	Meini prawf
5	Mynegiant clir iawn ac arddull ddarllenadwy iawn. Ychydig iawn o wallau iaith. Gwybodaeth berthnasol a dealltwriaeth dda o'r erthygl; heb fylchau arwyddocaol.
4	Mynegiant clir ac arddull ddealladwy. Nifer fach o wallau iaith. Gwybodaeth berthnasol a dealltwriaeth dda o'r erthygl, gyda rhai bylchau.
3	Mynegiant clir ac arddull ddealladwy ar y cyfan. Nifer o wallau iaith. Mae gwybodaeth a dealltwriaeth yr erthygl yn ddigonol, er bod nifer o bethau wedi'u gadael allan a nifer o wallau.
2	Mae'r mynegiant fel arfer yn glir ond weithiau yn aneglur. Nifer sylweddol o wallau iaith. Mae'r wybodaeth a'r ddealltwriaeth o'r erthygl yn ddigonol ar gyfer crynodeb elfennol, ond mae nifer o bethau wedi'u gadael allan a nifer o wallau.
1	Mae'r mynegiant yn aml yn anodd i'w ddeall. Arddull ddiffygiol. Gwallau iaith difrifol yn gyson. Mae'r wybodaeth yn annigonol at ddibenion crynhoi. Diffygion amlwg yn y ddealltwriaeth o'r erthygl.

[4] Cysill: https://www.cysgliad.com/cy/cysill/

Y sgôr gymedrig a'r sgôr canolrif ar gyfer y crynodebau oedd 4. Gofynnwyd i'r gwerthuswyr gywiro gwallau iaith cyffredin (megis gwallau treiglo a sillafu) ond i beidio â chywiro'r gystrawen.

5 DISGRIFIAD O'R OFFERYN CRYNHOI

Ail gam y project crynhoi hwn yw defnyddio set ddata'r corpws i oleuo datblygiad iteraidd a gwerthusiad offer crynhoi digidol. Mae prif ddulliau crynhoi testun yn cynnwys crynhoi ar sail alldynnu (*extraction-based summarisation*) a chrynhoi ar sail haniaethu (*abstraction-based summarisation*). Mae'r cyntaf yn alldynnu geiriau/ymadroddion penodol o'r testun i greu'r crynodeb, tra bo'r ail yn gweithio i ddarparu crynodebau wedi'u haralleirio (h.y. heb eu halldynnu'n uniongyrchol) o'r testun ffynhonnell. Mae alldynnu/haniaethu cynnwys, wrth ddefnyddio offer/dulliau crynhoi, yn dibynnu ar gywirdeb algorithmau awtomatig (sydd angen eu hyfforddi gan ddefnyddio setiau data safon aur wedi'u codio â llaw).

Mewn iaith brin ei hadnoddau fel y Gymraeg, lle na cheir llawer o lenyddiaeth ar grynhoi testun, mae cymhwyso technegau crynhoi o'r llenyddiaeth yn helpu i gael canlyniadau cychwynnol y gellir eu defnyddio i feincnodi perfformiad crynhowyr eraill ar y Gymraeg. Yn y project hwn rydym am ddatblygu cyfuniad o ddulliau alldynnol a haniaethu i grynhoi dogfennau unigol. Bydd y broses yn dechrau drwy weithredu a gwerthuso systemau gwaelodlin sylfaenol sydd yn cael eu defnyddio'n aml yn y llenyddiaeth fel meinclinau. Bydd y rhain yn cael eu dilyn gan fodelau crynhoi mwy cymhleth a blaengar yn ogystal â systemau hybrid fel penawdlinau (*toplines*).

5.1 Gwaelodlinau

Mae'r adrannau isod yn rhoi trosolwg o'r systemau crynhoi y bydd y project hwn yn canolbwyntio arnynt ar hyn o bryd yn ogystal ag yn ystod oes y project.

5.1.1 Crynhöwr Brawddeg Gyntaf

Yn hytrach na defnyddio teitl neu allweddeiriau dogfen, mae rhai crynhowyr yn tueddu i ddefnyddio brawddeg gyntaf erthygl i adnabod y pwnc sydd i gael ei grynhoi. Mae'r cyfiawnhad dros ddewis y frawddeg gyntaf fel un sy'n berthnasol i'r pwnc perthnasol yn seiliedig ar y gred fod y frawddeg gyntaf mewn llawer o achosion, yn enwedig mewn erthyglau papur newydd neu erthyglau a geir ar Wicipedia, yn tueddu i gynnwys gwybodaeth allweddol am gynnwys yr erthygl gyfan [6, 7, 8].

5.1.2 TextRank

Cyflwynwyd y dechneg grynhoi hon gan Rada Mihalcea a Paul Tarau [9]. Hwn oedd yr algorithm crynhoi testun awtomedig seiliedig ar graffau cyntaf i gael ei seilio ar ddefnydd syml o algorithm PageRank. Defnyddir PageRank gan Google Search i osod tudalennau gwe yn eu safle yn eu canlyniadau peiriant chwilio [10]. Mae TextRank yn defnyddio'r nodwedd hon i adnabod y brawddegau pwysicaf mewn erthygl.

5.1.3 LexRank

Yn debyg i TextRank, mae LexRank yn defnyddio algorithm seiliedig ar graffau ar gyfer crynhoi testun wedi'i awtomeiddio [11]. Seilir y dechneg ar y ffaith fod modd edrych ar glwstwr o ddogfennau fel rhwydwaith o frawddegau sy'n perthyn i'w gilydd. Mae rhai brawddegau yn debycach i'w gilydd ond efallai bod brawddegau eraill lle mai prin yw'r wybodaeth sy'n cael ei rhannu gyda gweddill y brawddegau. Yn debyg i TextRank, mae LexRank hefyd yn defnyddio algorithm PageRank i alldynnu'r allweddeiriau uchaf. Y gwahaniaeth allweddol rhwng y ddwy

waelodlin yw'r nodwedd bwysoli a ddefnyddir i aseinio pwysau i ymylon y graff. Tra bo TextRank yn cymryd yn syml fod pob pwysau yn bwysau uned ac yn cyfrifo safleoedd fel gweithrediad PageRank arferol, mae LexRank yn defnyddio graddau o debygrwydd rhwng geiriau ac ymadroddion ac yn cyfrifo pa mor ganolog yw'r brawddegau i aseinio pwysau [11].

5.2 Penawdlinau (*toplines*)

Wrth i'r project fynd rhagddo, byddwn yn datblygu crynhowyr mwy cymhleth ac yn gwerthuso eu perfformiad drwy gymharu canlyniadau crynhoi y dair gwaelodlin a enwir uchod. Diben y crynhowyr penawdlinau yw profi bod defnyddio technoleg berthnasol i iaith i grynhoi dogfennau Cymraeg yn gwella canlyniadau'r rhai a gynhyrchwyd gan y crynhowyr gwaelodlin.

5.2.1 Crynhöwr Cymraeg TF.IDF

Mae crynhöwr sy'n defnyddio TF.IDF (Text Frequency Inverse Document Frequency) yn gweithio ar ganfod geiriau sydd â'r gymhareb uchaf ar gyfer yr amlder geiriau hynny yn ddogfen sydd i'w chrynhoi o'i chymharu â pha mor aml maent yn digwydd yn y set lawn o ddogfennau sydd i'w crynhoi [12]. Ystadegyn rhifol syml yw TF.IDF sy'n adlewyrchu pwysigrwydd gair i ddogfen mewn casgliad o destunau neu gorpws ac fe'i defnyddir fel arfer fel ffactor bwysoli wrth adalw gwybodaeth, gan ei ddefnyddio felly i ganfod brawddegau pwysig mewn crynhoi alldynnol [13, 14].

Bydd y crynhöwr yn gweithio ar ganfod allweddeiriau a geiriau pwysig yn y dogfennau i'w crynhoi mewn ymgais i gynhyrchu crynodebau perthnasol. Nid peth newydd yw defnyddio TF.IDF yn y Gymraeg. Defnyddiodd Arthur et al. [15] rwydwaith gymdeithasol gan ddefnyddio geo-leoliadau Twitter i adnabod ardaloedd daearyddol cydgyffyrddol ac adnabod patrymau cyfathrebu o'u mewn a rhyngddynt. Yn yr un modd, byddwn yn defnyddio TF.IDF i adnabod brawddegau pwysig yn seiliedig ar batrymau a ganfuwyd rhwng y ddogfen a grynhowyd a'r corpws crynodebau.

5.2.2 Crynhöwr Cymraeg TF.IDF gyda Mewnblaniadau Geiriau Cymraeg

Er mwyn medru gwella'r mesur tebygrwydd rhwng brawddegau, bu i ni ddefnyddio nodweddion mewnblaniadau geiriau a raghyfforddwyd wedi'u cyfuno gyda'r nodweddion TF.IDF y soniwyd amdanynt eisoes. Ar gyfer hynny rydym yn defnyddio fectorau geiriau Cymraeg FastText [16]. Mae FastText yn estyniad o fodel word2vec [17] lle, yn hytrach na dysgu fectorau am eiriau yn uniongyrchol, mae FastText yn cynrychioli pob gair fel n-gram o nodau, sy'n helpu cipio ystyr geiriau byrrach a galluogi'r mewnblaniadau i ddeall olddodiaid a rhagddodiaid. Defnyddiodd Ezeani et al. [18] fodelau iaith oedd yn bod eisoes megis FastText Cymraeg ar gyfer dosbarthu rhannau ymadrodd Cymraeg a thagio semantig mewn tasgau lluosog. Byddwn yn ailadrodd yr arbrawf ond y tro hwn yn defnyddio mewnblaniadau geiriau Cymraeg a grëwyd gan Corcoran et al. [19] lle defnyddir word2vec a FastText i ddysgu mewnblaniadau geiriau Cymraeg yn awtomatig gan gymryd i ystyriaeth hynodion cystrawennol a morffolegol yr iaith. Byddwn yn adeiladu ar y ddwy ymdrech flaenorol hon ac yn harnesu'r modelau iaith i gyfoethogi perfformiad y crynhöwr TF.IDF yn 5.2.1.

5.3 Crynhowyr Cymraeg mwyaf blaengar

Cam olaf y project yw defnyddio'r technolegau crynhoi mwyaf blaengar i grynhoi dogfennau Cymraeg. Bydd hyn yn cynnwys adeiladu crynhowyr alldynnol a haniaethol gan ddefnyddio technegau dysgu peirianyddol rhwydweithiau

niwral dwfn neu'r hyn sy'n cael ei adnabod fel Dysgu Dwfn. Mae'r llenyddiaeth fwyaf blaengar ar grynhoi yn dangos symudiad mawr tuag at ddefnyddio dysgu dwfn i greu crynhowyr dan oruchwyliaeth a dioruchwyliaeth alldynnol a haniaethol gan ddefnyddio modelau dysgu dwfn megis CNN (Convolutional Neural Network), RNN (Recurrent Neural Network), ac LSTM (Long Short Term Memory) a nifer o rai eraill [20, 21, 22, 23]. Yn y project hwn byddwn yn cyfuno'r defnydd o fewnblaniadau geiriau Cymraeg y soniwyd amdanynt eisoes i geisio gwella'r canlyniadau a chreu systemau crynhoi Cymraeg sydd cystal â chrynhowyr blaengar eraill Saesneg ac Ewropeaidd.

6 GWERTHUSIAD

Defnyddir y crynodebau safon aur a grëwyd gan y crynhowyr dynol fel y disgrifiwyd yn Adran 4 i werthuso yn awtomatig unrhyw grynodebau system a gynhyrchwyd gan y modelau a ddatblygwyd yn Adran 5. Caiff y crynodebau system eu gwerthuso gan ddefnyddio metrigau ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [24]. Mae ROUGE yn gyfres o fetrigau a ddefnyddir i werthuso crynhoi awtomatig mewn prosesu iaith naturiol. Mae'r metrigau yn cymharu crynodeb sy'n cael ei gynhyrchu'n awtomatig yn erbyn crynodebau safon aur.

Fel lefel ychwanegol o werthuso, bydd sampl o grynodebau system a gynhyrchir gan ein model crynhoi sy'n perfformio orau (gw. Adran 5.2) yn cael eu gwerthuso â llaw gan siaradwyr Cymraeg brodorol er mwyn mesur ansawdd y crynodebau hynny. Bydd camau olaf y project hwn yn cynnwys datblygu rhyngwyneb defnyddiwr ar y we fydd ar gael yn agored ac a fydd yn gyfeillgar i'r defnyddiwr, a hefyd yn medru cael ei ddefnyddio gan bob grŵp oed. Bydd y system yn galluogi defnyddwyr i ddiffinio lefel y cywasgu (e.e. crynodeb o ddim mwy na 200 gair). Bydd y crynhöwr ar gael hefyd fel pecynnau Python cod agored i alluogi datblygwyr i weithio ar wella'r crynhowyr yn y dyfodol.

7 CASGLIADAU

Bydd y fersiwn o ACC gaiff ei ryddhau yn cyfrannu at yr offer awtomatig sydd ar gael yn y Gymraeg ac yn hwyluso gwaith y rhai sy'n ymwneud â pharatoi dogfennau, prawf-ddarllen, a (dan rai amgylchiadau) cyfieithu. Bydd yr offeryn hefyd yn galluogi gweithwyr proffesiynol i grynhoi dogfennau hir yn gyflym er mwyn eu cyflwyno'n effeithiol. Er enghraifft, bydd yr offeryn yn galluogi addysgwyr i addasu dogfennau hir i'w defnyddio yn yr ystafell ddosbarth. Rhagwelir hefyd y bydd yr offeryn o fudd i'r cyhoedd yn fwy cyffredinol, lle bydd efallai yn well ganddynt ddarllen crynodeb o wybodaeth gymhleth sy'n cael ei chyflwyno ar y rhyngrwyd neu sydd efallai yn ei chael hi'n anodd darllen fersiynau a gyfieithwyd o wybodaeth ar wefannau. I gael y diweddaraf am ddatblygiadau'r offeryn hwn, ewch i brif wefan y project ar: https://corcencc.org/acc/

DIOLCHIADAU

Ariannwyd yr ymchwil hwn gan Lywodraeth Cymru dan y Grant Crynhoi Testun Cymraeg Awtomatig. Rydym yn ddiolchgar i Jason Evans, Wicipediwr Cenedlaethol Llyfrgell Genedlaethol Cymru, am ei gyngor cychwynnol.

CYFEIRIADAU

[1] Jeremy Evas a Daniel Cunliffe. 2016. Behavioural Economics and Minority Language e-Services—The Case of Welsh. Yn Durham, M. a Morris, J. (eds), Sociolinguistics in Wales. Palgrave Macmillan. Llundain. 61-91.

[2] Cen Williams. 1999. Cymraeg Clir: Canllawiau Iaith. Bangor: Cyngor Gwynedd, Bwrdd yr Iaith Gymraeg a Chanolfan Bedwyr.

[3] Llywodraeth Cymru. 2017. Cymraeg 2050 - A million Welsh speakers. Adalwyd o https://gov.wales/sites/default/files/publications/2018-12/cymraeg-2050-welsh-language-strategy.pdf

[4] Patrick Carlin a Diarmait Mac Giolla Chríost. 2016. A standard for language? Policy, territory, and constitutionality in a devolving Wales. Yn Durham, M. a Morris, J. (eds), Sociolinguistics in Wales. Palgrave Macmillan. Llundain. 93-119.

[5] Uned Technolegau Iaith Prifysgol Bangor. 2021. Cysgliad: Help i ysgrifennu yn Gymraeg. Adalwyd o https://www.cysgliad.com/cy/

[6] Dragomir R. Radev, Hongyan Jing, Małgorzata Styś a Daniel Tam. 2004. Centroid-based Summarization of Multiple Documents. Information Processing and Management, 40, 919–938.

[7] M. Fattah a F. Ren. 2008. Automatic Text Summarization. Yn Proceedings of World Academy of Science, 27, World Academy of Science, 192–195.

[8] Jen-Yuan Yeh, Hao-Ren Ke a Wei-Pang Yang. 2008. iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network. Expert Systems with Applications, 35(3), 1,451–1,462.

[9] Rada Mihalcea a Paul Tarau. 2004. TextRank: Bringing Order into Text. Yn Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Sbaen, 404-411.

[10] Sergey Brin a Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7), 107-17.

[11] Erkan, G. a Radev, D. R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research, 22, 457–479.

[12] Gerard Salton a Michael J. McGill. 1986. Introduction to modern information retrieval. McGraw Hill. Efrog Newydd.

[13] Hajime Mochizuki a Manabu Okumura. 2000. A Comparison of Summarization Methods based on Taskbased Evaluation. Yn Proceedings of the 2nd International Conference on Language Resources and Evaluation. Athens, Groeg, 404-411.

[14] C. G. Wolf, S. R. Alpert, J. G. Vergo, L. Kozakov a Y. Doganata. 2004. Summarizing Technical Support Documents for Search: Expert and User Studies. IBM Systems Journal, 43(3), 564–586.

[15] Rudy Arthur a Hywel T. P. Williams. 2019. The human geography of Twitter: Quantifying regional identity and inter-region communication in England and Wales. PloS one, 14 (4), e0214466.

[16] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou a Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. Rhag. 12, 2016. arXiv:1612.03651.

[17] Tomas Mikolov, Kai Chen, Greg Corrado a Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Ion. 16, 2013. arXiv:1301.3781.

[18] 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). Florence, Yr Eidal, 270-280.

[19] Padraig Corcoran, Geraint Palmer, Laura Arman, Dawn Knight, ac Irena Spasić. 2021. Creating Welsh language word embeddings. Applied Sciences, 11(15), 6896.

[20] Shengli Song, Haitao Huang a Tongxiao Ruan. 2019. Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools and Applications, 78(1), 857-875.

[21] Nadhem Zmandar, Abhishek Singh, Mahmoud El-Haj a Paul Rayson. 2021. Joint abstractive and extractive method for long financial document summarization. Yn Proceedings of the 3rd Financial Narrative Processing Workshop. Prifysgol Caerhirfryn, Caerhirfryn, Lloegr, 99-105.

[22] Mahmoud El-Haj, Nadhem Zmandar, Paul Rayson, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, George Giannakopoulos, Aris Kosmopoulos, Blanca Carbajo-Coronado a Antonio Moreno-Sandoval. 2021. The Financial Narrative Summarisation Shared Task FNS 2021. Yn Proceedings of the 3rd Financial Narrative Processing Workshop. Prifysgol Caerhirfryn, Caerhirfryn, Lloegr, 120-125.

[23] P. G. Magdum a Sheetal Rathi. 2021. A Survey on Deep Learning-Based Automatic Text Summarization Models. Advances in Artificial Intelligence and Data Engineering. Springer. Singapore. 377-392.

[24] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. Yn Text summarization branches out. Association for Computational Linguistics. Barcelona, Sbaen, 74–81.

Yr hyn sydd ei angen er mwyn cael Adnabod Lleferydd Awtomatig gweithredol ar gyfer y Gernyweg

PREBEN VANGBERG

Prifysgol Bangor

LEENA SARAH FARHAT

Prifysgol Bangor

Mae'r ymchwil hwn yn edrych ar heriau ac atebion posibl er mwyn adeiladu system adnabod lleferydd weithredol ar gyfer y Gernyweg. Mae'r Gernyweg yn iaith Frythonig a siaredid gynt yng Nghernyw, Lloegr, ond a fu farw yn y ddeunawfed ganrif. Mae wedi cael ei hadfywio yn ystod y degawdau diwethaf, ac mae angen cynyddol am system adnabod lleferydd sy'n medru adnabod a thrawsgrifio lleferydd Cernyweg. Mae'r ymchwil hwn yn nodi dwy broblem sylweddol wrth adeiladu system adnabod lleferydd awtomatig ar gyfer y Gernyweg: prinder data hyfforddi a ffonoleg unigryw yr iaith. Mae'r astudiaeth ragarweiniol hon yn darparu amryw o atebion posib i'r materion hyn, gan gynnwys defnyddio dysgu dim-siot, hyfforddi model ar orgraff Gernyweg addasedig, a defnyddio ffonemau fel cam yn y canol yn y broses drawsgrifio. Mae'r bennod hon hefyd yn archwilio'r angen am set ddata awdio wedi'i thrawsgrifio ac sydd ar gael yn gyhoeddus, ac yn awgrymu y gallai Common Voice Mozilla fod yn llwyfan addas i letya set ddata o'r fath. Mae'r ymchwil yn cloi gan gynnig y gellid defnyddio detholiad o'r methodolegau hyn i greu system adnabod lleferydd awtomatig fyddai'n gweithio ar gyfer y Gernyweg, ond yn cydnabod yr angen am fwy o waith yn y maes hwn.

Allweddeiriau: Adnabod Lleferydd Awtomatig, Cernyweg, Orgraff y Gernyweg, Technoleg Lleferydd

1 RHAGARWEINIAD

Mae'r bennod hon yn cyflwyno'r heriau a'r dulliau posib o ddatblygu system Adnabod Lleferydd Awtomatig ar gyfer y Gernyweg heb ddata addas ar gyfer y dasg.

1.1 Y Gernyweg

Mae'r Gernyweg yn iaith Frythonig sy'n cael ei siarad yng Nghernyw. Mae'n perthyn i'r un teulu a'r Llydaweg a'r Gymraeg ac yn perthyn o bell i Aeleg yr Alban, Manaweg a Gwyddeleg. Arferai Cernyweg fod yn brif iaith Cernyw, ond dechreuodd edwino ac ildio'i lle i'r Saesneg o'r 13g ganrif ymlaen. Erbyn canol y ddeunawfed ganrif yr oedd wedi peidio i bob pwrpas a bod yn iaith fyw, ond mae wedi gweld adfywiad sylweddol yn y degawdau diweddar, diolch i unigolion a mudiadau brwdfrydig sydd wedi ymrwymo i gadw a hybu diwylliant Cernyw. Cafwyd problemau niferus wrth geisio adfer y Gernyweg, yn eu plith diffyg safon cydnabyddedig i ysgrifennu'r iaith. Cynigiwyd sawl orgraff wahanol, ond ni ddatryswyd hyn tan 2008, pan gafwyd cytundeb cyffredinol o blaid y Ffurf Ysgrifenedig Safonol (Standard Written Form neu SWF), ond gyda'r amrywiadau eraill oedd mewn bri megis Kernewek Kemmyn, Cernyweg Unedig, Cernyweg Tuduraidd a Cernyweg Diweddar Wedi'i Adfywio hefyd yn cael eu cydnabod [1]. Nododd 563 o bobl eu bod yn siarad Cernyweg yng nghyfrifiad y DU yn 2021 [2], gydag ychydig filoedd o ddysgwyr hefyd yn dysgu'r iaith. Mae'r pwll bychan hwn o siaradwyr yn ei gwneud hi'n anodd casglu digon o ddata lleferydd i hyfforddi modelau adnabod lleferydd yn ddigonol i ddatblygu system adnabod lleferydd ar gyfer yr iaith. Ar waethaf y rhwystrau hyn, mae'r gymuned Gernyweg wedi bod yn weithgar yn creu cynnwys ar gyfer eu hiaith. Mae mentrau fel radio a theledu Cernyweg yn galluogi Cernywiaid i ddefnyddio'r iaith mewn sefyllfaoedd bob dydd, gan wella'r amgylchedd ieithyddol. Er nad ydynt yn uniongyrchol yn ychwanegu at y pwll o siaradwyr brodorol, mae'r nifer sylweddol o ddysgwyr Cernyweg hefyd yn chwarae rhan bwysig yn adfywio'r iaith.

Gall y dysgwyr niferus hefyd newid ynganiad y Gernyweg. Oherwydd fod y myfyrwyr yn dod o amrywiol gefndiroedd ieithyddol, gallant ddod ag ynganiadau newydd neu amrywiadau ar seiniau sy'n bodoli'n barod gyda nhw. Mae'r duedd hon, sy'n cael ei galw'n 'ymyrraeth,' yn elfen arferol o ryngweithiad ieithoedd a gall arwain at dwf iaith dros amser. Oherwydd y nifer fechan o siaradwyr Cernyweg, efallai fod effaith dysgwyr arni yn fwy amlwg. Nid yw'r effaith hon fodd bynnag bob amser yn niweidiol; gall gyfrannu at hyblygrwydd a gwydnwch iaith.

Mae'r gymuned Gernyweg yn gweithio'n barhaus i drwsio problemau ynganu ac i sicrhau unffurfiaeth yn normau'r iaith. Yn gyffredinol, mae adfywiad y Gernyweg yn dyst rhyfeddol i bŵer parhaus iaith a phenderfyniad unigolion a chymunedau sy'n ymroi i ddiogelu eu hetifeddiaeth ieithyddol. Er bod rhwystrau yn bodoli, yn arbennig o ran casglu corpws lleferydd mawr ar gyfer systemau adnabod lleferydd, mae'r gymuned Gernyweg yn gweithio i fynd i'r afael â'r anawsterau hyn ac yn gwneud cynnydd mawr yn adfywio'r iaith unigryw a gwerthfawr hon.

1.2 Adnabod Lleferydd Awtomatig

Caiff modelau adnabod lleferydd awtomatig eu hyfforddi ar setiau data anferth o leferydd wedi'i labelu. Gall fod cyn heriol, os nad yn amhosibl, casglu digon o ddata wedi'i labelu i hyfforddi model adnabod lleferydd effeithiol mewn ieithoedd prin eu hadnoddau, megis y Gernyweg. Cafwyd cynnydd yn y diddordeb mewn adeiladu systemau adnabod lleferydd ar gyfer ieithoedd prin eu hadnoddau yn y blynyddoedd diweddar. Mae hyn yn rhannol oherwydd datblygiad dysgu dwfn, sydd wedi galluogi hyfforddi modelau adnabod lleferydd gyda llai o ddata. Yn ychwanegol, cafwyd mwy o gydnabyddiaeth o'r angen i warchod ieithoedd, a gall adnabod lleferydd chwarae rhan mewn diogelu ieithoedd prin eu hadnoddau. Mae'r bennod hon yn ceisio trafod beth sydd ei angen i alluogi adnabod lleferydd defnyddiol ar gyfer y Gernyweg.

2 DATA HYFFORDDI AR GYFER ADNABOD LLEFERYDD A THECHNOLEGAU LLEFERYDD ERAILL

Fel arfer, mae angen dwy gydran allweddol ar gyfer creu system adnabod lleferydd, model acwstig i drosi lleferydd yn destun, a model iaith i gywiro gwallau yn y testun hwnnw. Caiff y rhain eu hyfforddi mewn ffyrdd gwahanol ac maent angen data gwahanol.

2.1 Data hyfforddi ar gyfer modelau acwstig

Gall yr holl ffurfiau amrywiol ar y Gernyweg, yn enwedig fersiynau hŷn, effeithio ar ynganiad. Bydd siaradwyr efallai yn glynu wrth safonau ynganu'r ffurf sydd fwyaf cyfarwydd iddynt hwy, gan olygu bod gwahaniaethau yn y ffordd mae gwahanol eiriau ac ymadroddion yn cael eu ynganu. Mae hyn yn ffenomenon arferol mewn unrhyw iaith sydd ag amrywiaeth o ffurfiau, ac nid yw o reidrwydd yn peryglu unoliaeth yr iaith. Wrth hyfforddi modelau adnabod lleferydd, mae'n hanfodol dangos amrywiaeth eang o ynganiadau Cernyweg i'r model, yn enwedig rhai sy'n gysylltiedig gyda gwahanol amrywiadau orgraffyddol. Bydd hyn yn cynorthwyo cyffredinoliad a pherfformiad y modelau lle mae siaradwyr yn defnyddio ynganiadau ansafonol. Yr unig ddata sy'n hollol benodol i'r cymhwysiad adnabod lleferydd yw lleferydd wedi'i drawsgrifio. Defnyddir hyn i hyfforddi'r rhan acwstig o'r system adnabod lleferydd gyfan. Er bod synthesis lleferydd hefyd yn defnyddio lleferydd wedi'i drawsgrifio, ceir yn aml wahaniaeth yn yr ansawdd. Mae ar synthesis lleferydd angen awdio o safon da heb sŵn cefndir fel bod y llais yn gliriach. Gall adnabod lleferydd ar y llaw arall elwa ar awdio swnllyd gan ei fod yn gwneud y modelau yn fwy cadarn. Mae gan academyddion fynediad at recordiadau o ddarlleniadau a chyfweliadau drwy gyfrwng y Gernyweg (gwaith a wnaeth gan Szczepankiewicz [3]). Fodd bynnag, er ei bod hi'n bosibl fod peth o'r data hwn wedi'i drawsgrifio, nid yw eto wedi'i segmentu i glipiau fesul brawddeg i'w wneud yn ddefnyddiol ar gyfer hyfforddi modelau acwstig. Mae

recordiadau eraill ar gael ac wedi'u trawsgrifio o siaradwyr yn defnyddio'r iaith, ond mae'r rhan fwyaf o'r rhain dan drwydded, a heb fod ar gael i'w defnyddio i ddatblygu adnabod lleferydd.

2.1.1 Common Voice Mozilla

Mae Common Voice Mozilla [4] yn gorpws lleferydd wedi'i dorfoli a ddatblygwyd ac sy'n cael ei gynnal gan Mozilla. Mae ganddo setiau data ar gyfer 114 iaith ac mae nifer y setiau data yn cynyddu. Ar hyn o bryd nid oes gan Common Voice set ddata weithredol ar gyfer y Gernyweg, ond mae ymdrechion i gael set ddata Gernyweg ar y gweill. Y cam cyntaf i lawnsio set ddata ar Common Voice yw lleoleiddio rhyngwyneb Common Voice yn yr iaith darged. Ar 30 Hydref 2023, roedd 56% o'r rhyngwyneb wedi'i leoleiddio. Mae Common Voice hefyd angen set o frawddegau dan drwydded Creative Commons Sero (CC0) i ddefnyddwyr eu recordio. Byddai cael set ddata Common Voice Cernyweg ar Common Voice yn gyfle arbennig i'r iaith ac adnabod lleferydd Cernyweg gan y byddai'n annog siaradwyr i gyfrannu data ar gyfer datblygu technolegau iaith mewn dull hawdd i'w ddefnyddio drwy ddulliau torfoli.

2.2 Data hyfforddi ar gyfer modelau iaith

Mae modelau Iaith gan amlaf yn defnyddio testun plaen fel data hyfforddi. Yn ffodus, ceir tipyn go lew o destunau Cernyweg ar y rhyngrwyd y gellir eu defnyddio at y diben hwn, ond nid yw'n hawdd dod o hyd iddynt na'u defnyddio. Yn aml, roedd yn rhaid i ni ddibynnu ar Archif y Rhyngrwyd i gael digon o destun Cernyweg y gallem ei ddefnyddio, ac roedd tipyn o waith glanhau arno oherwydd natur y data. Nid yw hyn yn amhosibl, ac mae wedi cael ei wneud ar gyfer ieithoedd eraill (e.e. mae wedi cael ei wneud ar gyfer y Llydaweg yn y gorffennol [5]), ond nid yw'n rhywbeth sydd wedi'i wneud eto ar gyfer y Gernyweg.

2.2.1 Labelu Testun Cernyweg

Fel y soniwyd uchod, yn ddiweddar safonwyd orgraff y Gernyweg yn un ffurf safonol, sy'n fuddiol i unffurfiaeth a hygyrchedd yr iaith. Rhaid nodi fodd bynnag fod yr amrywiadau cynharach o sillafu Cernyweg sy'n dal i fodoli yn cael eu defnyddio gan rai siaradwyr a dysgwyr. O ganlyniad, mae'n hanfodol ceisio cael cydbwysedd rhwng glynu wrth y safon SWF a chaniatáu brawddegau mewn amrywiadau eraill. Bydd y dull cynhwysol hwn yn helpu dal gafael yn nyfnder a bywiogrwydd yr iaith gan gynnal derbyn y safon SWF ar yr un pryd. Mae angen mwy o ymchwil ar effaith y gwahanol orgraffau hyn ar effeithiolrwydd y modelau iaith. Efallai y byddai cael modelau iaith ar wahân yn targedu'r gwahanol orgraffau yn rhoi gwell canlyniadau nag un model cyfun. Er mwyn ymchwilio i hyn, mae angen i setiau data gynnwys gwybodaeth yn nodi pa orgraff a ddefnyddiwyd. Gyda digon o destun wedi'i dagio felly, gallai fod yn bosib awtomeiddio'r dosbarthiad i ryw raddau drwy ddefnyddio offer megis Tawa [6].

3 METHODOLEG

Treialwyd sawl methodoleg ar gyfer y project. Dewiswyd y dulliau hyn yn unol â chefndir ieithyddol y Gernyweg yn ogystal ag ystyried y prinder data.

3.1 Dim-siot

Mae dysgu dim-siot yn galluogi modelau i wneud tasgau heb gael eu hyfforddi'n benodol arnyn nhw. Yn wahanol i ddysgu safonol dan oruchwyliaeth, sy'n hyfforddi modelau ar gasgliad o enghreifftiau wedi'u labelu, gall modelau dysgu dim-siot gyffredinoli i dasgau a data newydd heb hyfforddiant ychwanegol. Mae arbrofion blaenorol wedi

dysgu y gall defnyddio modelau acwstig a gynlluniwyd ar gyfer ieithoedd a thafodieithoedd sy'n perthyn yn agos, ar y cyd â modelau iaith a gynlluniwyd yn benodol, gynhyrchu canlyniadau go lew [7]. Felly gwnaed rhai arbrofion rhagarweiniol gan ddefnyddio fframwaith Wav2Vec2 [8] a chan ddefnyddio modelau a hyfforddwyd ar gyfer gwahanol ieithoedd. Gellir gweld y modelau hyn yn Nhabl 2. Yn ychwanegol at wneud profion gan ddefnyddio gwahanol fodelau acwstig, profwyd gwahanol gyfluniadau o destun i hyfforddi ein modelau iaith. Daw'r testunau a ddefnyddiwyd oddi wrth Gyngor Cernyw (a elwir yn Corpws Kernewek yng ngweddill y bennod) ac allforiad o'r Wikipedia Cernyweg (gweler https://dumps.wikimedia.org/kwwiki/latest/). Profwyd pob un ar ei ben ei hun yn ogystal ag fel model cyfun.

3.2 Hyfforddi gan ddefnyddio orgraff wedi'i haddasu

Yr iaith sy'n perthyn agosaf at y Gymraeg yw'r Gernyweg. Mae nifer helaeth o eiriau ac ymadroddion yn tarddu o'r Frythoneg a'u rhagflaenai, ac mae rhai newidiadau orgraffyddol rheolaidd yn ein galluogi i olrhain y berthynas rhyngddynt. Er enghraifft, yn Gymraeg, caiff y seiniau /v/ ac /f/ eu hysgrifennu fel ⟨f⟩ ac ⟨ff⟩, ond mewn Cernyweg cânt eu hysgrifennu gyda ⟨v⟩ ac ⟨f⟩ fel mewn Saesneg. Mae gan y Gernyweg ffonoleg gymhleth, gyda rhai nodweddion yn ei gosod ar wahân i'r ieithoedd Brythonig eraill. Mae cymathiad ⟨d⟩ mewn Hen Gernyweg i ⟨s⟩ mewn Cernyweg Canol, er enghraifft, *dad* (tad mewn Cymraeg) mewn Hen Gernyweg a *tas* mewn Cernyweg Canol, yn un o nodweddion hynotaf y Gernyweg. Credir fod y newid hwn wedi digwydd tua'r unfed ganrif ar ddeg [9].

Un o nodweddion eraill ffonoleg y Gernyweg yw taflodiad /d/ i /dʒ/ o flaen llafariaid blaen, fel mewn *dzadn* (dant mewn Cymraeg) mewn Cernyweg Canol a /dʒayn/ mewn Cernyweg Modern. Credir fod y shifft yma wedi digwydd o gwmpas y drydedd ganrif ar ddeg. Mae gan y Gernyweg hefyd nifer o ffonemau llafariaid gwahanol, gan gynnwys y llafariaid byr /I/ ac /oe/ a'r llafariaid hir /I:/ ac /oe:/. Credir fod y llafariaid hyn wedi esblygu o ddeuseineiddio i ac u Hen Gernyweg o flaen cytseiniaid trwynol.

Gan fod y model adnabod lleferydd dim-siot Cymraeg yn gwneud camgymeriadau orgraffyddol amlwg, crëwyd ffwythiant Python syml i drosi gwahaniaethau orgraffyddol syml mewn Cymraeg i safon SWF y Gernyweg. Yna mireiniwyd y modelau Wav2Vec2 gan ddefnyddio'r data Cymraeg hyn wedi'u 'Cernywegeiddio' mewn ymdrech i gynorthwyo'r modelau adnabod lleferydd Cymraeg i drawsgrifio'r ffonemau hyn yn gywir.

3.3 Defnyddio ffonemau fel cam yn y canol

Mae'r Wyddor Ffonetig Ryngwladol (IPA) [10] yn ffordd annibynnol ar iaith o drawsgrifio lleferydd. Prawf arall oedd defnyddio IPA fel cam yn y canol. Gan fod IPA yn annibynnol ar iaith, golyga hyn fod modd trawsgrifio i IPA gan ddefnyddio modelau wedi'u hyfforddi i'r diben hwnnw ar gyfer ieithoedd eraill, ac yna drosi'r IPA hwnnw i orgraff y Gernyweg.

3.3.1 Sain i ffonem

Y cam cyntaf yw trosi'r awdio i IPA. Mae nifer o ffyrdd o wneud hyn. Mae Allosaurus [11] yn ddarn o feddalwedd a gynlluniwyd at y diben hwn. Gellir cael hyd i sawl model gwahanol ar HuggingFace (gweler https://huggingface.co/). Defnyddiwyd y model Wav2Vec2 gan Phy [12] fel prawf cysyniad.

3.3.2 Ffonem i graffem

Mae'r cam nesaf ychydig yn fwy cymhleth. Nid yw'r cam hwn yn annhebyg i dasg gyfieithu arferol; fodd bynnag mae'n seiliedig ar nodau unigol yn hytrach na geiriau. Mae yna ddwy ffordd o brototeipio'r systemau yn gyflym; roedd y gyntaf yn system seiliedig ar reolau a'r llall yn fater o hyfforddi model dilyniant-i-ddilyniant.

Dull seiliedig ar reolau: Cryfder y dull seiliedig ar reolau yw ei bod hi'n eithaf hawdd taflu rhywbeth at ei gilydd a chael canlyniadau go lew. Y gwendid fodd bynnag yw nad ydyw yn 'ddeallus' iawn ac mae'n eithaf hawdd mynd i drybini gan ei bod yn anodd cymryd i ystyriaeth y cyd-destun ble mae'r ffonem yn digwydd.

Yn anffodus nid oedd yr IPA o'r model yn cynnwys unrhyw wybodaeth am hyd y ffonemau, ac roedd yn anodd cymryd i ystyriaeth eginiad a hyd llafariaid. Mae rhai llafariaid yn amwys hefyd. Er enghraifft, yn dibynnu ar yr orgraff, gall /E/ gael ei ysgrifennu fel ⟨e⟩ neu ⟨eu⟩, a gellir ysgrifennu /i:/ fel ⟨i⟩ neu ⟨y⟩.

Er mwyn goresgyn hyn dewiswyd y Tawa Toolkit [6] gyda'i nodweddion trawsnewidiol i ddatrys rhai o'r materion hyn drwy greu set o reolau trawsffurfio i drosi cytseiniaid unigol yn gytseiniaid dwbl a datrys llafariaid amwys. Hyfforddwyd y model iaith a ddefnyddiwyd ar gyfer hyn ar destun gan Gyngor Cernyw. Ni chafodd y rheolau hyn eu profi na'u hoptimeiddio ond eu penderfynu mewn dull ad-hoc yn seiliedig ar yr hyn yr amcangyfrifwyd fyddai'n gwella'r canlyniadau.

Defnyddio modelau dilyniant-i-ddilyniant: Roedd rhannau o'r geiriadur ynganu yng ngeiriadur ar-lein yr Akademi Kernewek yn hygyrch i ni (gweler https://www.cornishdictionary.org.uk/). Y rhagdybiaeth oedd y gellid defnyddio hyn i greu model dilyniant-i-ddilyniant. Profwyd sawl math o Gofau Hir Tymor Byr (LSTMs), Trawsffurfwyr, a modelau eraill. Fodd bynnag, nid oedd digon o ddata i hyfforddi'r modelau hyn yn effeithiol, felly rhoddwyd y gorau i'r dull hwn.

4 CANLYNIADAU

Yn gyffredinol, roedd y canlyniadau yn addawol ar gyfer y gwaith rhagarweiniol hwn. Mae'r arbrofion hyn yn dangos bod adnabod lleferydd awtomatig ar gyfer y Gernyweg yn ddichonadwy ond yn dal yn bell o fod yn ddigon da hyd yn hyn.

4.1 Data profi a ddefnyddiwyd

At ddibenion profi, dim ond un darn llafar Cernyweg o tua 15 munud o hyd oedd ar gael i ni. Holltwyd y recordiad hwn â llaw yn segmentau ar gyfer ei brofi. Oherwydd ystyriaethau preifatrwydd ni ellir darparu manylion pellach am y recordiad. Roedd y set brofi hon yn heriol oherwydd nid darn o destun yn cael ei ddarllen ydoedd, ond yn hytrach berfformiad wedi'i ddramateiddio. Arweiniodd hyn at ynganiad wedi'i or-ddramateiddio mewn rhai brawddegau.

4.2 Modelau Iaith

Y nod cyntaf oedd archwilio effeithiolrwydd gwahanol ffurfweddiadau ein modelau iaith. Optimeiddiwyd yr holl fodelau iaith a'u profi gan ddefnyddio'r data profi a'r model techiaith/wav2vec2-xlsr-ft-en-cy. Profwyd cyfanswm o dri model fel sy'n cael ei ddangos yn Nhabl 1. Er i'r model Corpws Kernewek berfformio'n well na'r lleill, nid yw hyn yn debyg o fod yn ystadegol arwyddocaol, am fod cyn lleied o achosion profi yn y set ddata.

Tabl 1: Trosolwg o wahanol gyfraddau gwallau yn y tri model iaith (LM) o ran canran

Corpws Testun	WER[1]	WER gyda LM	CER[2] gyda LM
Corpws Kernewek	95.05	78.37	43.54
Wikipedia	95.05	83.26	44.93
Y cyfan	95.05	83.42	45.24

4.3 Adnabod Lleferydd Awtomatig Dim-siot

Ar gyfer y Gernyweg, un prif anfantais o ddysgu dim-siot yw y gallai fod yn sensitif i anawsterau orgraffyddol. Mae gan y Gernyweg orgraff gymhleth gyda sawl nodwedd arbennig, gan gynnwys y defnydd o nodau diacritig a chymathiad *d* i *s*. Mae nifer o fodel dysgu dim-siot heb eu dysgu i ddelio gyda phroblemau sillafu, a gall hyn arwain at gamgymeriadau. Mae gan y Gymraeg, un o'r ieithoedd sy'n perthyn agosaf i'r Gernyweg, sillafiad mwy rheolaidd. Fodd bynnag, mae'n wahanol iawn i'r Gernyweg. Fel y gwelir yn Nhabl 2, ni wnaeth yr ieithoedd berfformio yn arwyddocaol wahanol. Fodd bynnag, mae'r canlyniadau yn wahanol pan gaiff model iaith ei gynnwys, ac mae'r model techiaith/wav2vec2-xlsr-ft-en-cy yn rhagori cryn dipyn ar y lleill.

4.4 Hyfforddi gan ddefnyddio orgraff wedi'i haddasu

Hyfforddwyd ystod o fodelau Wav2Vec2 gan ddefnyddio amrywiaeth o hyper-baramedrau gwahanol. Fodd bynnag, ar gyfer pob ymgais, methodd yr hyfforddi â dysgu unrhyw beth o werth. Y canlyniadau terfynol oedd WER o 95.58% a CER o 86.67%, wrth ddefnyddio model iaith. Sylwch mai'r rheswm ei fod yn llai na 100% yw ei fod weithiau yn allbynnu bylchau gwyn gan amlaf, sy'n golygu ei fod weithiau yn gywir. Efallai mai'r rheswm iddo fethu dysgu unrhyw beth oedd bod y data wedi torri felly mae'r canlyniadau yn amhendant. I gloi, er y gellid cael sefyllfa lle byddai modd hyfforddi model gwell, nid yw hynny'n sicr o bell ffordd.

Tabl 2: Trosolwg o'r gwahanol gyfraddau gwallau ar gyfer y modelau dim-siot yn ôl canran, wedi'u drefnu yn ôl CER.

Iaith	Model	WER heb LM	WER	CER
CY	techiaith/wav2vec2-xlsr-ft-cy	96.47	86.41	59.26
XX	voidful/wav2vec2-xlsr-multilingual-56	95.82	84.46	55.90
BR	DrishtiSharma/wav2vec2-large-xls-r-300m-br-d2	96.90	91.74	53.01
EN	jonatasgrosman/wav2vec2-large-xlsr-53-english	96.90	82.83	51.95
EN-CY	techiaith/wav2vec2-xlsr-ft-en-cy	95.05	78.37	43.54

4.5 Defnyddio ffonemau fel cam yn y canol

Canfu ein profion fod gan y system hon WER o 100% a CER o 48.63% heb unrhyw fodelau iaith gan ddefnyddio dull hynod o syml o amnewid ar sail rheolau. Gan nad oedd yr un o'r camau yn y broses hon wedi'i optimeiddio, ac nad oedd cefnogaeth i gytseiniaid dwbl, nac ychwaith unrhyw fodel iaith, mae hyn yn eithaf addawol. Dangosodd y profion hefyd fod cynnwys y rheolau trawsnewid Tawa yn gwella'r canlyniadau, ond nid oes ffigurau pendant ar y gwelliant ar gael eto.

[1] Cyfradd Gwallau Geiriau
[2] Cyfradd Gwallau Nodau

5 CASGLIADAU A GWAITH PELLACH

Er bod adnabod lleferydd awtomatig ar gyfer y Gernyweg yn ddichonadwy ar hyn o bryd, mae'n bell o fod cystal ag y gallai fod, ac mae angen llawer o waith eto i wella ar hyn.

Mae data lleferydd yn hanfodol ar gyfer gwaith adnabod lleferydd am sawl rheswm. Mae'n galluogi systemau adnabod lleferydd awtomatig i ddysgu priodoleddau acwstig y Gernyweg, gan gynnwys y gwahanol seiniau a sut maent yn cael eu cyfuno. Mae data lleferydd hefyd yn cynorthwyo systemau adnabod lleferydd i ddysgu'r berthynas ystadegol rhwng seiniau a geiriau, sy'n angenrheidiol i drosi sain yn eiriau mewn modd cywir. Yn ychwanegol at hyfforddi systemau adnabod lleferydd, gellir defnyddio data lleferydd hefyd i ddatblygu technolegau iaith eraill, megis testun i leferydd a systemau cyfieithu peirianyddol. Bydd cael mwy o ddata lleferydd ar gyfer y Gernyweg yn cynnal datblygiad technolegau iaith newydd a gwell. Byddai cael set ddata Common Voice Cernyweg ar Common Voice yn gam mawr ymlaen ar gyfer y Gernyweg ac adnabod lleferydd Cernyweg. Byddai'n ei gwneud hi'n hawdd i siaradwyr gyfrannu data i ddatblygu technolegau iaith.

Mae gwaith pellach, unwaith bod yna set ddata y gellir ei defnyddio fel sylfaen ar gyfer technolegau iaith, yn cynnwys datblygu modelau adnabod lleferydd sy'n fwy cadarn i ddelio ag amrywiadau'r iaith. Mae gan y Gernyweg sawl amrywiad, felly mae'n bwysig datblygu modelau adnabod lleferydd sy'n medru eu trawsgrifio i gyd a mewn modd dibynadwy. Un dull fyddai defnyddio corpora hyfforddi fyddai'n cynnwys data o nifer o amrywiadau gwahanol.

CYFEIRIADAU

[1] Albert Bock a Benjamin Bruc. 2008. An Outline of the Standard Written Form of Cornish. Retrieved from (PDF). https://kernowek.net/Specification_Final_Version.pdf.

[2] Office for National Statistics. 2022. Language, England and Wales - office for national statistics. Adalwyd o https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language/bulletins/languageenglandandwales/census2021

[3] P. Szczepankiewicz. 2023. Linguistic variation in revived Cornish. Research grant 2022/45/N/HS2/00869. National Science Centre, Adam Mickiewicz University. Adalwyd o https://projekty.ncn.gov.pl/index.php?projekt_id=557852

[4] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers a Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. Maw. 5, 2020. arXiv:1912.06670.

[5] Preben Vangberg a Leena Farhat. 2023. Speech-to-text for Breton. Yn Celtic student conference, 2023.

[6] William John Teahan. 2018. A Compression-Based Toolkit for Modelling and Processing Natural Language Text. Information 9(12), 294. https://doi.org/10.3390/info9120294

[7] Preben Vangberg a Leena Farhat. 2023. Devisa: Exploring transfer learning in an interdialectal setting for Romansch. Yn the XIX International Conference on Minority Languages, 2023.

[8] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed a Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. Hyd. 22, 2020. arXiv: 2006.11477.

[9] Talat Chaudhri. 2007. Studies in the Consonantal System of Cornish. Thesis PhD, Adran y Gymraeg ac Astudiaethau Celtaidd, Prifysgol Aberystwyth.

[10] I. P. Association. 1999. Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet. Cambridge University Press. Cambridge.

[11] Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W Black a Florian Metze . 2020. Universal phone recognition with a multilingual allophone system. Yn IEEE international conference on acoustics, speech and signal processing (ICASSP) 2020. IEEE, 2020, Barcelona, Sbaen, 8249–8253.

[12] Vitou Phy. 2022. Automatic phoneme recognition on TIMIT dataset with wav2vec 2.0. Adalwyd o https://huggingface.co/vitouphy/wav2vec2-xls-r-300m-timit-phoneme

Sut i weithio gyda ieithoedd llai eu hadnoddau a modelau iaith mawr

COLIN JARVIS

OpenAI

Mae'r bennod hon yn archwilio sut mae modelau iaith mawr (LLMs), fel ChatGPT, yn trin ieithoedd â data hyfforddi cyfyngedig. Darperir disgrifiad o sut mae LLMs yn cael eu hyfforddi a sut y gellir eu hoptimeiddio. Gan ddefnyddio cywiro gwallau Islandeg a Thestun-i-Leferydd (TTS) Cymraeg fel astudiaethau achos, mae'r bennod hon yn amlinellu'r cyfyngiadau sydd ynghlwm wrth y modelau hyn, strategaethau ar gyfer eu mireinio, a dulliau ymarferol o roi'r strategaethau hyn ar waith mewn ieithoedd sydd ag adnoddau cyfyngedig mewn sefyllfaoedd go iawn.

Allweddeiriau: Modelau Iaith Mawr, Optimeiddio, Data Hyfforddi, Islandeg, Cymraeg, Ieithoedd Llai eu Hadnoddau

1 CYFLWYNIAD

Mae modelau iaith mawr (LLMs) wedi cyffroi dychymyg llawer o bobl ers iddyn nhw ddod i olwg y cyhoedd pan lansiwyd ChatGPT. Un o agweddau mwyaf diddorol eu poblogrwydd newydd yw priodweddau datblygiadol modelau iaith mawr (megis GPT gan OpenAI, Claude gan Anthropic, Google Gemini, Llama 2 ac ati) gyda ieithoedd lle nad oedd llawer o ddata yn y setiau hyfforddi. Pwrpas y bennod hon yw ffocysu ar yr ieithoedd llai eu hadnoddau hynny, a sut mae modd eu hoptimeiddio ar gyfer tasgau ymarferol.

Mae'r bennod hon yn manylu ar agweddau ymarferol defnyddio ieithoedd llai eu hadnoddau gyda modelau iaith mawr mewn tair prif ffrwd:

- sut i gyflwyno iaith newydd i fodel iaith mawr yn ystod ei broses hyfforddi, a'r cyfyngiadau sy'n codi wrth ddefnyddio modelau sylfaen ar gyfer tasgau ieithyddol
- crynhoi'r technegau mwyaf poblogaidd presennol ar gyfer pensaernïo achosion defnydd amlieithog ac optimeiddio modelau iaith mawr sylfaen,[1] gan gynnwys peiriannu promptiau, cynhyrchu adalw estynedig (retrieval-augmented generation: RAG) a mireinio
- sut i ddefnyddio'r technegau hyn i bensaernïo cymwysiadau modelau iaith mawr ar gyfer ieithoedd llai eu hadnoddau, a sut i'w hoptimeiddio i'w rhoi ar waith wrth gynhyrchu cymwysiadau.

Erbyn diwedd y bennod hon byddwch yn deall yn well gyfyngiadau'r modelau hyn, y technegau ar gyfer eu tiwnio, a sut i gymhwyso hyn at ieithoedd llai eu hadnoddau yn y byd go iawn.

2 SUT MAE MODELAU IAITH MAWR YN GWEITHIO, A SUT I'W HYFFORDDI

Fe ddechreuwn ni drwy grynhoi, mor syml ag y gallwn, sut mae modelau iaith mawr yn cael eu hyfforddi, er mwyn rhoi cyd-destun i ble yn y broses maen nhw'n dysgu gwybodaeth newydd, a sut mae hynny'n dod â chyfyngiadau i berfformiad amlieithog.

Caiff model iaith mawr ei hyfforddi mewn dau gam cychwynnol, gydag thrydydd cam dewisol, fel y dangosir yn Nhabl 1.

[1] Mae model sylfaen yn derm sy'n disgrifio'r gwahanol fodelau iaith mawr sy'n cael eu cynnig ar y farchnad fel modelau 'sylfaen', y mae modd naill ai eu defnyddio'n uniongyrchol neu eu mireinio ymhellach ar gyfer achosion defnydd penodol. Enghreifftiau o hyn yw modelau GPT-3.5-turbo a GPT-4 OpenAI, modelau Gemini Google, modelau Claude Anthropic, Llama gan Meta, a llawer eraill.

Tabl 1: Camau hyfforddi LLM

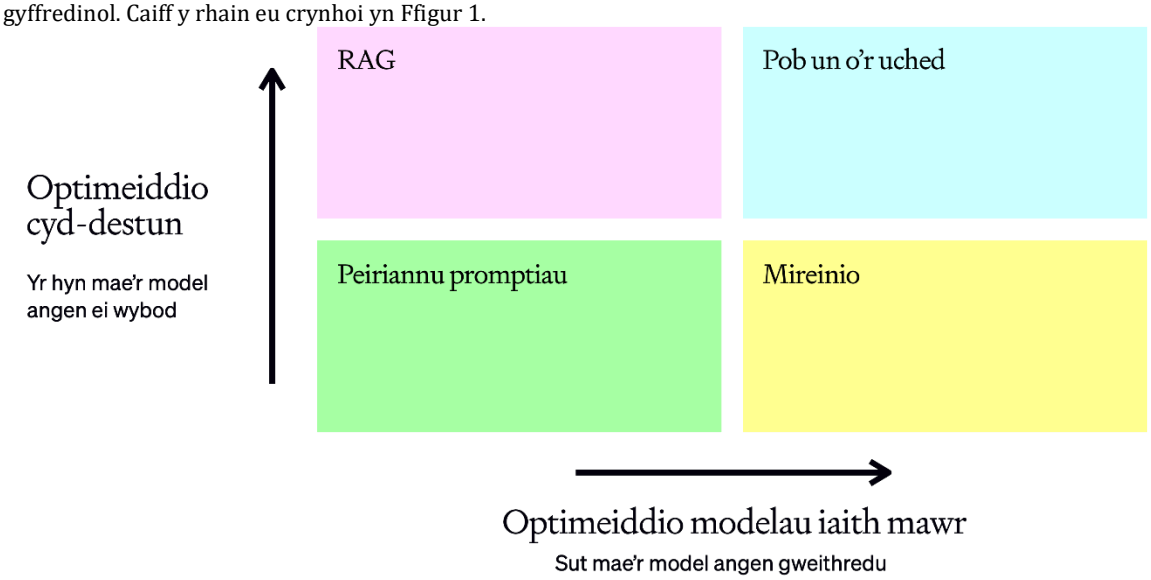
Cam	Disgrifiad	Effaith iaith
Rhag-hyfforddi	Y cam cyntaf yw rhag-hyfforddi, lle caiff meintiau anferthol o ddata eu cyflwyno i'r model. Y ffactor bwysicaf yma yw maint, lle rydych chi'n cyflwyno corpws llawn y wybodaeth rydych eisiau i'r model gael gwybodaeth amdani, ac mae'r model yn dysgu rhagweld yn effeithiol y tocyn nesaf o ystyried y tocynnau blaenorol. Ar ddiwedd y broses hon mae'r model yn gwybod sut i ragweld y tocyn nesaf yn dda iawn, ond nid oes unrhyw gyfyngiadau na gwerthoedd wedi'u adeiladu i mewn iddo, ac nid yw'n medru dilyn cyfarwyddiadau na rhyngweithio'n effeithiol gyda bodau dynol heb eu hyfforddi.	Er mwyn cael perfformiad amlieithog cryf mae angen i'r model gael gymaint o ddata yn eich iaith ag sy'n bosib yn y cam rhag-hyfforddi. Mae ansawdd yn llai perthnasol yma, gan y bydd digon o faint yn dysgu iddo'r strwythurau mwyaf cyffredin a ddefnyddir yn eich iaith. Yn anffodus mae rhag-hyfforddi hefyd yn ddrud iawn, ac felly fel arfer nid yw hwn yn opsiwn sydd ar gael i'r defnyddiwr modelau iaith mawr arferol.
Ôl-hyfforddi	Mae ôl-hyfforddi yn cynnwys sawl cam sy'n ymwneud â thechnegau megis mireinio dan oruchwyliaeth, modelu gwobrwyol neu ddysgu atgyfnerthol. Bwriad y cam yma yw dysgu'r model pa allbynnau sydd orau gan ddefnyddwyr, sut i ddilyn cyfarwyddiadau/sgwrsio gyda phobl a chyflwyno'r weithred o wrthod er mwyn blocio cynnwys anniogel neu anghyfreithlon. Gan ddibynnu ar y dull a ddefnyddir, y flaenoriaeth yma yw nifer llai o enghreifftiau hyfforddi ansawdd uchel. Dyma lle gallwch chi diwnio gallu'r model i berfformio tasgau – er enghraifft, darparu nifer o enghreifftiau o siarad Islandeg fel ei fod yn dysgu pa strwythur gramadegol sydd orau.	Mae ôl-hyfforddi yn cynnig cyfle gwych i diwnio dealltwriaeth y model o sut i ddefnyddio iaith benodol. Er enghraifft, gallech roi cynnig ar gywiro nodweddion gramadegol rydych wedi sylwi bod y model yn eu cynhyrchu, neu ychwanegu'r data perthnasol ar gyfer cyfuniad parth/iaith penodol. Byddai modd cyrraedd y ddau nod yma drwy gyflwyno enghreifftiau hyfforddi wedi'u tocio'n dda yn gynharach yn y broses ôl-hyfforddi. Mae ôl-hyfforddi yn llai drud ac yn cymryd llai o ddata na rhag-hyfforddi, felly mae'n ddewis mwy poblogaidd er ei fod yn dal y tu allan i gyrraedd y rhan fwyaf o ddatblygwyr sydd heb yr adnoddau i ôl-hyfforddi modelau iaith mawr eu hunain.
Mireinio	Mae mireinio yn gam dewisol lle mae modd mireinio model sylfaen sydd ar gael ar y farchnad neu drwy drwydded cod agored gyda'ch data chi eich hun. Dyma'r opsiwn isaf ei gost, ond oni bai eich bod yn gwneud mireinio ar raddfa fawr sy'n debycach i ôl-hyfforddi, does dim modd effeithiol i chi ychwanegu gwybodaeth newydd i'r model. Fel arfer, bydd eich mireinio yn addasu pwysau'r model, gan ei wella ar gyfer tasg neu barth cyfyng penodol.	Mireinio yw'r dull mwyaf cost effeithiol o gynyddu perfformiad iaith, ond mae'n gyfyngedig i wella gallu'r model mewn ieithoedd y mae ganddo eisoes wybodaeth ohonynt. Os nad oedd gan fodel ddiogon o ddata yn eich iaith cyn dechrau, eich unig opsiwn i wella hyn yw ôl-hyfforddi neu rag-hyfforddi.

I grynhoi, mae rhag-hyfforddi yn cychwyn gyda llawer iawn o ddata i ddysgu ieithoedd i'r model, gall ôl-hyfforddi ychwanegu setiau bach arbenigol at hyn, yn ogystal â thiwnio ei ddefnydd o'r ieithoedd y mae wedi'u dysgu, ac mae mireinio yn gweithio orau ar gyfer cynyddu perfformiad gyda ieithoedd sydd eisoes yn hysbys iddo.

Rydym wedi gweld, hyd yn oed ar gyfer ieithoedd llai eu hadnoddau fel Cymraeg, Gwyddeleg ac Islandeg, fod digon o wybodaeth gan nifer o fodelau sylfaen presennol i wneud mireinio a thechnegau optimeiddio eraill yn effeithiol. Mae gweddill y bennod hon yn cymryd fod gan y model rydych chi'n ei ddefnyddio rywfaint o wybodaeth o'ch iaith, a bydd yn eich helpu i wasgu gymaint o berfformiad ag y mae modd gan ddefnyddio'r technegau tiwnio modelau iaith gorau.

3 OPTIMEIDDIO CYMWYSIADAU MODELAU IAITH MAWR

Cyn i ni sôn am sut i optimeiddio modelau iaith mawr ar gyfer eu defnyddio gyda ieithoedd llai eu hadnoddau yn benodol, mae'n werth mynd dros yr arferion gorau cyfredol ar gyfer optimeiddio modelau iaith mawr yn gyffredinol. Caiff y rhain eu crynhoi yn Ffigur 1.



Ffigur 1: Matrics Optimeiddio

Yma, ceir rhai egwyddorion i'w hystyried:

- **nid yw optimeiddio yn llinol:** dechreuwch bob tro gyda gwerthusiad o berfformiad cyfredol eich model. I wybod ble i optimeiddio nesaf, mae angen i chi arunigo'r broblem i'r hyn mae'r model angen ei wybod, neu sut mae angen iddo weithredu. Hynny sy'n dweud wrthych beth i'w wneud nesaf
- **dechreuwch gyda phrompt a gwerthusiad:** y cam cyntaf yw ysgrifennu prompt, a cheisio datrys eich tasg ar rai enghreifftiau lle rydych chi'n gwybod beth yw'r canlyniad cywir. Mesurwch sut gafodd y model y canlyniad anghywir, dyma yw eich gwaelodlin
- **er mwyn cael cyd-destun, defnyddiwch RAG:** os mai cyd-destun yw'r hyn sydd ei angen ar y model iaith mawr, megis data perchnogol, e.e. ymchwil neu erthyglau papur newydd, neu hyd yn oed dim ond rhai enghreifftiau perthnasol iddo gyfeirio atynt, yna RAG yw'r cam nesaf cywir. Bydd hyn fel arfer yn cynnwys

rhyw fath o chwiliad, fydd yn adalw cyd-destun perthnasol a'i osod yn y prompt i'w ddefnyddio gan y model adeg tynnu casgliadau

- **er mwyn cysondeb, defnyddiwch mireinio: m**ae mireinio yn ddefnyddiol lle mae'r model yn dilyn cyfarwyddiadau yn anghyson. Casglwch set o enghreifftiau hyfforddi wedi'u tocio'n dda (byddai hyd yn oed 100 yn ddigon i ddechrau arni), a mireiniwch fodel, yna defnyddiwch y model hwnnw i ail-werthuso i weld a wnaeth y cysondeb wella. Os wnaeth e, ond eich bod angen rhagor o welliant, parhewch i ychwanegu enghreifftiau
- **mae optimeiddio yn gweithio'r ddwy ffordd:** gall y broses hon fod yn faith, felly peidiwch â cholli gobaith os ydych chi'n rhoi cynnig ar rywbeth a bod y perfformiad yn gwaethygu. Mae angen i chi fod yn systematig drwy newid newidynnau cyfyngedig gyda phob iteriad, a gwerthuso bob tro.

Y peth olaf i'w bwysleisio yw y gall cyd-destun a chysondeb ill dau fod yn broblemus yn aml. Yn ffodus, dydi RAG a mireinio ddim yn cau ei gilydd allan, mae gan y ddau rywbeth i'w gyfrannu, ac mae modd eu defnyddio gyda'i gilydd pan fo angen er mwyn cael y perfformiad gorau o ystyried cost a chyfnod cyn ymateb.

Mae unrhyw daith tuag at optimeiddio yn ceisio cydbwyso tair ffactor, sef **cywirdeb**, **cyfnod cyn ymateb** a **chost**. Mae bob amser yn werth cadw'r rhain mewn cof wrth ystyried eich cam optimeiddio nesaf, a'r strategaeth fwyaf cyffredin yw cynyddu cywirdeb hyd at y pwynt lle rydych chi'n cael perfformiad derbyniol, lle gallwch symud mwy tuag at y cyfnod cyn ymateb a lleihau cost (defnyddio llai o docynnau, modelau llai o faint neu lai o alwadau API).

Felly, i grynhoi, dylem bob amser ddechau gyda pheiriannu promptiau, gwerthuso'r allbwn, ac yna benderfynu sut i fynd ymlaen yn dibynnu ar p'un a yw'r datrysiad angen rhagor o gyd-destun, cysondeb, neu'r ddau. Gyda hyn mewn cof, gadewch i ni symud ymlaen a chymhwyso hyn at yr her ymarferol o optimeiddio ar gyfer ieithoedd llai eu hadnoddau.

4 ASTUDIAETHAU ACHOS AR GYFER IEITHOEDD LLAI EU HADNODDAU

I ddangos y dulliau optimeiddio hyn yn gweithio, rydym wedi defnyddio dau achos defnydd yn ymdrin â moddau testun ac awdio:

1. **Cywiriadau Islandeg:** rydym yn defnyddio'r Corpws Gwallau Islandeg [1] i brofi a fedrwn ni wella perfformiad modelau OpenAI GPT 3.5 a GPT-4 yn cywiro Islandeg
2. **Testun-i-Leferydd Cymraeg:** rydyn ni wedyn yn newid i awdio, gan ddefnyddio model testun-i-leferydd OpenAI gan fewnbynnu awdios cyfeirnodol [2] i wella hyd y gallwn berfformiad awdio sy'n cael gynhyrchu yn Gymraeg.

4.1 Cywiriadau Islandeg

Mae'r Corpws Gwallau Islandeg yn cynnwys cyfuniadau o frawddeg Islandeg gyda gwallau, ynghyd â'r fersiwn wedi'i chywiro o'r frawddeg honno. Byddwn yn defnyddio model GPT-4 i geisio datrys y dasg hon, ac yna yn cymhwyso gwahanol dechnegau optimeiddio i weld sut y gallwn wella perfformiad y model.

4.1.1 Arbrawf

Fe ddechreusom ni drwy fformatio'r data i'w fewnbynnu i fodelau GPT. Roedd hyn yn golygu gwneud prompt `system` gyda chyfarwyddiadau i'r model eu dilyn, ac yna ddarparu'r frawddeg oedd i'w chywiro fel neges `user`. Mae ateb y cynorthwyydd yn cynnwys ymgais y model i gyfieithu. Gellir gweld enghraifft unigol yn Nhabl 2.

Tabl 2: Enghraifft o ddata wedi'i fewnbynnu i fodelau GPT

system	user	cynorthwyydd
The following sentences contain Icelandic sentences which may include errors. Please correct these errors using as few word changes as possible.	Sörvistölur eru nær hálsi og skartgripir kvenna á brjótsti.	Sörvistölur eru nær hálsi og skartgripir kvenna á brjósti.

Ar gyfer gwerthuso fe ddefnyddion ni ddau fetrig oddi-ar-y-silff i gyfrifo perfformiad perthynol pob model:

- pellter golygu: pellter golygu Levenshtein rhwng y cywiriad ei hun a'r rhagfynegiad
- BLEU [3]: sgôr BLEU wedi'i ddefnyddio i fesur ansawdd y cywiriad a ragwelwyd o'i gymharu â'r cywiriad cyfeirnodol.

Byddwn yn rhoi cynnig ar y dulliau canlynol ac yn mesur y sgorau gwerthuso ar gyfer pob un:

- GPT-4 heb unrhyw enghreifftiau (Dim-siot)
- GPT-4 gyda 3 enghraifft (Rhai-siots)
- GPT-3.5 wedi'i fireinio gyda 1,000 enghraifft
- GPT-4 wedi'i fireinio gyda 1,000 enghraifft
- GPT-4 wedi'i fireinio gyda 1,000 enghraifft + 3 enghraifft debyg o RAG

Defnyddiodd ein piblinell RAG 1,000 o enghreifftiau o gywiriadau oedd wedi'u dal allan, gafod eu mewnblannu gan ddefnyddio ada-v2-embeddings OpenAI a'u storio mewn cronfa ddata fectorau qdrant. Cânt eu hadalw gan ddefnyddio tebygolrwydd cosin i enghraifft y mewnbwn.

4.1.2 Canlyniadau

Dangosir y canlyniadau llawn yn Ffigur 2. Roedd y canlyniadau yn ddiddorol, ac fe welsom ni fod mireinio yn gweithio'n dda ar gyfer y dasg hon, ond bod RAG mewn gwirionedd yn niweidio perfformiad. Y negeseuon allweddol o'r dadansoddiad oedd:

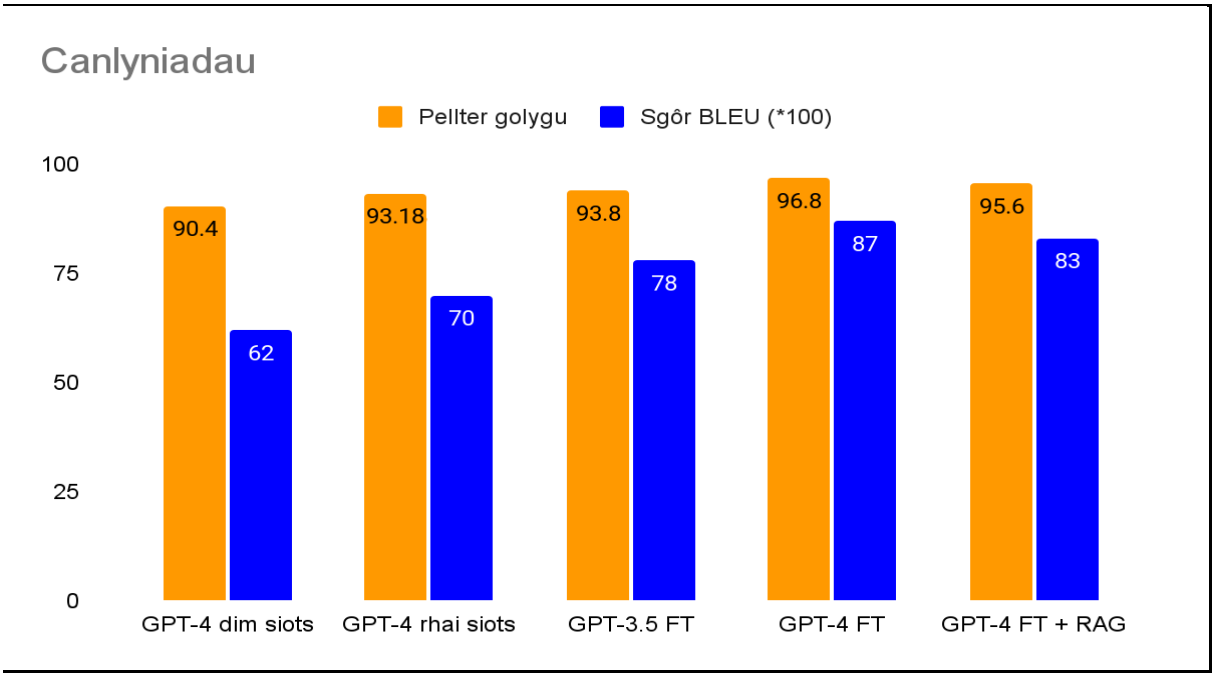
- roedd GPT-4 gyda rhai-siots yn sylweddol well na GPT-4 dim-siot, gan wella'r sgôr BLEU o 8 pwynt
- roedd perfformiodd GPT-3.5 wedi'i fireinio gyda 1,000 enghraifft yn well na GPT-4 gyda rhai-siots, gan gynnig opsiwn llawer mwy cost-effeithiol
- GPT-4 FT oedd y gorau, yn perfformio'n well hyd yn oed na GPT-4 FT + enghreifftiau.

Un rhybudd ynghylch yr uchod yw pan fyddwch chi'n ychwanegu RAG at fodel sydd wedi'i fireinio, byddem ni'n awgrymu eich bod yn ailhyfforddi'r model a fireiniwyd fel ei fod yn cael ei hyfforddi ar enghreifftiau sydd â RAG ynddyn nhw. Mae perfformiad wedi'i fireinio yn gwyro lle bynnag mae'r enghreifftiau hyfforddi yn sylweddol wahanol o'r cynhyrchiad, fel yn yr achos yma.

4.1.3 Casgliad

Ein dull gorau oedd GPT-4 wedi'i fireinio gyda 1,000 enghraifft, gan fireinio dealltwriaeth y model o sut i gywiro Islandeg i gael sgôr BLEU uchel o 87.

Profasom hefyd fod ein dull optimeiddio yn gweithio – dechreusom gyda rhai-siots, cadarnhau fod hynny'n gwella pethau, yna gwella eto drwy fireinio. Roedd ychwanegu RAG yn gwaethygu pethau, a allai awgrymu nad oes angen ychwanegu rhagor o gynnwys ar y cam o dynnu casgliadau, ac mewn gwirionedd mai optimeiddio'r model ar gyfer y dasg gywiro yw'r peth pwysicaf lle mae angen optimeiddio.



Ffigur 2: Canlyniadau arbrawf cywiriadau Islandeg

4.2 Testun-i-Leferydd Cymraeg

Yn awr fe drown ein sylw at un o foddau eraill modelau iaith mawr, lle byddwn yn edrych ar optimeiddio model Testun-i-Leferydd (TTS) OpenAI i gymryd i mewn destun Cymraeg a chynhyrchu lleferydd Cymraeg wedi'i acennu a'i ynganu yn gywir. Mae'r adran hon yn defnyddio fersiwn arbrofol o'r model sy'n medru derbyn awdio cyfeirnodol wedi'i fewnbynnu fel y llais i'w ddefnyddio ar gyfer ei gynhyrchu – mae hyn yn caniatáu gwella'r cynhyrchiad, er enghraifft gyda siaradwr sydd ag acen Gymraeg fel y cyfeirnod.

Ein nod yw cynhyrchu awdio gydag acen Gymraeg dda o gael testun mewnbwn Cymraeg.

4.2.1 Arbrawf

I optimeiddio awdio does dim gymaint o liferi gyda ni ag sydd ar gyfer testun, ond mae rhai opsiynau gennym ni o hyd, sef:

- **Peiriannu promptiau**
  - ailysgrifennu'r testun i ddarparu sillafiadau ffonetig lle mae geiriau yn cael eu camynganu.
  - defnyddio atalnodi i yrru rhythm llefaru h.y. nodau coll geiriau ar gyfer oediadau hirach, dileu atalnodau llawn neu ddefnyddio ffurfiau hirach geiriau pan fyddant yn rhy fyr yn yr awdio sy'n cael ei gynhyrchu
  - 'hacio promptiau' drwy ailysgrifennu'r testun i ddefnyddio mwy o dermau o'r ardal benodol rydych am i'r acen ei chynrychioli.
- **Awdio cyfeirnodol**
  - defnyddio awdio cyfeirnodol o ansawdd da gyda lleiafswm arteffactau a rhythm lleferydd cytbwys ar draws y sampl

- defnyddio siaradwr brodorol o'r iaith neu'r acen fel yr awdio cyfeirnodol
- defnyddio awdio cyfeirnodol o siaradwr heb fod yn un brodorol yn siarad yr iaith neu'r acen rydych chi am ei chynhyrchu.

- **Paramedrau model**
  - gall addasu cyflymder yr awdio sy'n cael ei gynhyrchu wella ansawdd ar gyfer cyfuniad siaradwr/iaith.

Ar gyfer yr arbrawf hwn byddwn yn canolbwyntio ar yr awdio cyfeirnodol a pharamedrau'r model, ac yn asesu ansawdd y cynhyrchu ar bob cam. Mae barnu'r ansawdd yn weithred eithaf oddrychol, felly fe wnawn ni greu rwbrig syml er mwyn cael rhywfaint o gysondeb:

- **Acen (1-5):** pa mor dda yw'r acen Gymraeg.
- **Ansawdd lleferydd (1-5):** ydi e'n swnio fel Cymraeg naturiol, ydi pob gair yn cael ei lefaru, ac a oedd yno unrhyw rithweledigaethau.
- **Ansawdd awdio (1-5):** ydi'r awdio sy'n cael ei gynhyrchu yn swnio'n naturiol, oes ganddo unrhyw arteffactau, niwlogrwydd neu ddangosyddion eraill o ansawdd gwael.

Y gwahanol awdios cyfeirnodol a ddefnyddiwyd:

- **TTS diofyn:** cynhyrchu gydag un o'r lleisiau TTS diofyn
- **Siaradwyr Cymraeg:** defnyddio awdios cyfeirnodol o siaradwr brodorol benywaidd a gwrywaidd
- **Siaradwr Cymraeg heb fod yn frodorol:** defnyddio awdio cyfeirnodol ohonof i fy hun, sydd ddim yn siaradwr Cymraeg brodorol.

4.2.2 Canlyniadau

4.2.2.1 TTS diofyn

Fel y gwelir yn Nhabl 3, mae'r TTS diofyn yn perfformio yn ganolig-wael ar y Gymraeg:

- mae'r acen yn Americanaidd iawn, ond mae'n medru ynganu *dd, f* a *ch* yn gywir
- mae ansawdd y lleferydd yn eithaf da, ond mae geiriau fel *mynadd* mor fyr fel eu bod nhw'n cael eu colli, ac mae un o'r llythrennau *i* yn mynd ar goll hefyd
- mae ansawdd yr awdio yn dda.

Tabl 3: Gwerthusiad o'r canlyniadau a gafwyd wrth ddefnyddio'r TTS diofyn fel awdio cyfeirnodol

Siaradwr	Acen	Ansawdd y lleferydd	Ansawdd yr awdio
alloy	2	3	5

4.2.2.2 Siaradwyr Cymraeg

Fel y gwelir yn Nhabl 4, fe gawsom ni ganlyniadau cymysg-gadarnhaol wrth ddefnyddio siaradwyr Cymraeg yn y sampl awdio. Y prif ganfyddiadau yma yw:

- yn gyffredinol roedd ansawdd yr awdio yn wael ar gyfer y llais benywaidd am fod yr awdio cyfeirnodol yn wael, felly er bod yr acen a chynnwys y lleferydd yn eithaf da, ar y cyfan roedd yn waeth na'r un gwrywaidd
- daeth yr awdio gwrywaidd allan yn ansawdd uwch, gyda llawer mwy o rythm y Gymraeg yn yr acen a'r lleferydd. Fe wnaeth ansawdd yr awdio cyfeirnodol effeithio'n andwyol ar hwn hefyd, ond roedd yr argraff gyffredinol yn dda.

Tabl 4: Gwerthusiad o'r canlyniadau a gafwyd wrth ddefnyddio siaradwyr Cymraeg fel awdio cyfeirnodol

Siaradwr	Acen	Ansawdd y lleferydd	Ansawdd yr awdio
Gwrywaidd	5	5	3
Benywaidd	4	3	1

4.2.2.3 Siaradwr Cymraeg heb fod yn frodorol

Fel y gwelir yn Nhabl 5, cafwyd canlyniadau cymysg hefyd o redeg fy llais fy hun, ond daeth â rhai darganfyddiadau diddorol yn ei sgil:

- roedd Cymraeg fy awdio cyfeirnodol yn araf ac elfennol, ac fe welwyd hyn yn y recordiadau cyflymder 1.0 a 0.8, lle roedd yr acen yn eithaf da ond nad oedd yn swnio fel siaradwr hyderus (rhaid dweud fod hynny'n wir hefyd)
- fodd bynnag, roedd y recordiad sain 1.2 yn swnio'n llawer mwy naturiol, gan fod y rhythm yn llawer nes at siaradwr Cymraeg brodorol
- yn anffodus roedd ansawdd yr awdio yn wael am fod yr awdio cyfeirnodol o ansawdd gwael, ond er gwaethaf hyn daeth geiriau fel *mynadd* drwodd yn iawn a doedd yna ddim drychiolaethau.

Tabl 5: Gwerthusiad o'r canlyniadau a gafwyd wrth ddefnyddio siaradwyr Cymraeg heb fod yn frodorol fel awdio cyfeirnodol

Siaradwr	Cyflymder	Acen	Ansawdd y lleferydd	Ansawdd yr awdio
Fi fy hun	1.0	3	4	3
	1.2	4	4	3
	0.8	2	4	3

4.2.2.4 Casgliad

Y pethau i sylwi arnynt yma yw:

- **mae angen tiwnio'r modelau sylfaenol:** ni fydd modelau iaith testun-i-leferydd allan o'r bocs fel arfer yn ddigon da ar gyfer ieithoedd llai eu hadnoddau oherwydd y duedd tuag at Saesneg gydag acen Americanaidd yn llawer o'u data hyfforddi
- **mae awdio cyfeirnodol yn allweddol:** gallwch ddylanwadu'n drwm ar y llais drwy ddefnyddio awdio cyfeirnodol cryf a model sy'n galluogi cydweddu lleisiau. Gyda'r dull hwn fe welsom:
    - i'n hacen orau ddod wrth ddefnyddio awdio ffynhonnell oedd yn siaradwr Cymraeg brodorol
    - ein hail orau oedd siaradwr Cymraeg heb fod yn frodorol yn siarad Cymraeg fel ffynhonnell, ond yn cael ein model i gynyddu cyflymder y cynhyrchu fel bod y rhythm yn cydweddu i rythm siaradwr brodorol.
- **paramedrau a all helpu:** helpodd ein paramedrau sain ni i gael cynhyrchiad mwy naturiol yn yr achos wnaethom ni ei ddefnyddio. Dydi hyn ddim bob amser yn gweithio, ond lle mae gennych gyfeirnod heb fod yn frodorol sy'n siarad yn araf, gall wneud y cynhyrchiad yn fwy naturiol.

Yn awr mae gennych rai o'r dulliau y medrwch eu defnyddio i optimeiddio testun-i-leferydd ar gyfer ieithoedd llai eu hadnoddau. Gallwch brofi'r rhain gydag Elevenlabs, gTTS neu unrhyw un o'r rhaglenni TTS eraill sydd ar y farchnad ac sy'n cynnig y gallu i gydweddu lleisiau.

5 CASGLIADAU

Gall defnyddio modelau iaith mawr sylfaen ar gyfer ieithoedd llai eu hadnoddau weithio'n dda iawn os yw gwybodaeth o'r iaith honno eisoes ar gael ynddynt, a gallwn hefyd ddefnyddio dulliau fel peiriannu promptiau, RAG a mireinio i ychwanegu at yr hyn sydd eisoes yno. Yn y bennod hon rydyn ni'n rhoi hynny ar waith, gan optimeiddio GPT-3.5 a GPT-4 i wella eu perfformiad ar Gywiriadau Islandeg, a rhai technegau mwy newydd i greu awdio sy'n siarad Cymraeg.

Ar gyfer y bobl hynny sydd eisiau hyd yn oed mwy o berfformiad o'r modelau iaith mawr hyn, mae rhag-hyfforddi ac ôl-hyfforddi ar gael o hyd, fel y soniwyd uchod. Ar gyfer y rhai sydd â'r adnoddau i wneud hynny, dyma'r ffordd fwyaf effeithiol i fynd o'i chwmpas hi, ond hefyd y ffordd fwyaf dwys o ran cost ac adnoddau.

Gobeithio fod hyn wedi rhoi trosolwg i chi o sut i ddefnyddio modelau iaith mawr gydag achosion defnydd ieithoedd lIai eu hadnoddau, a bod gennych yr hyn rydych ei angen i fynd allan a chreu llawer iawn mwy o achosion defnydd er mwyn i'r byd weld mwy o ffrwyth deallusrwydd artiffisial cynhyrchiol mewn amgylcheddau a moddau lleol. Byddwn yn croesawu ac yn gwerthfawrogi unrhyw adborth, ac rwy'n edrych ymlaen i weld beth fyddwch chi'n ei adeiladu.

REFERENCES

[1] Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir a Xindan Xu. 2021. Icelandic error corpus (IceEC) version 1.1. Adalwyd o https://github.com/antonkarl/iceErrorCorpus

[2] Uned Technolegau Iaith Prifysgol Bangor. 2023. Corpws Talentau Llais. Adalwyd o https://git.techiaith.bangor.ac.uk/data-porth-technolegau-iaith/corpws-talentau-llais

[3] Kishore Papineni, Salim Roukos, Todd Ward, a Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. Yn Pierre Isabelle, et al., (eds), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. ACL, Philadelphia, Pennsylvania, UDA, 311–318.

Gwerthusiadau Cyntaf o GPT OpenAI ar gyfer y Gymraeg

GRUFFUDD PRYS

Prifysgol Bangor

DEWI BRYN JONES

Prifysgol Bangor

Ym mis Tachwedd 2022, lansiwyd ChatGPT gan OpenAI, sef gwasanaeth sgyrsiol a ddaeth yn boblogaidd iawn mewn amser byr iawn. Defnyddir modelau iaith mawr GPT o'i fewn ac yn fuan iawn, darganfu defnyddwyr y gwasanaeth bod modd iddynt sgwrsio â'r ap yn Gymraeg. Annisgwyl hefyd oedd bod safon y Gymraeg yn syndod o dda, ond nid oedd yn berffaith chwaith, ac o ganlyniad i hynny teimlwyd y byddai'n werthfawr mesur gallu'r modelau gyda'r Gymraeg ac adnabod y mannau lle'r oedd angen ei wella. Cyflwynir yn y bennod hon, felly, y gyfres gyntaf o werthusiadau a gynlluniwyd i wella ein dealltwriaeth o hyd a lled y problemau ieithyddol Cymraeg gyda modelau GPT-3.5 a GPT-4. Ein bwriad oedd craffu yn fanwl ar wahanol broblemau ieithyddol penodol fel ehangder geirfa, y gallu i ymateb yn gadarnhaol neu'n negyddol i gwestiynau, a'r ddealltwriaeth o ramadeg cywir, yn ogystal â agweddau diwylliannol unigryw fel y gallu i adnabod iaith anweddus a'r gallu i ymdrin ag enwau lleoedd dwyieithog Cymru yn briodol. Yn ogystal, gwerthuswyd gallu'r modelau GPT i gyflawni tasgau cyfieithu peirianyddol parth benodol. Awgryma ein canlyniadau y gallai modelau iaith mawr fod o gymorth mawr i gyfieithwyr, ond y gallai cost defnyddio'r modelau mwyaf soffistigedig fel GPT-4 fod yn ormodol. O ganlyniad i'n gwerthusiadau ar y canlyniad o fireinio modelau rhatach fel GPT-3.5, rhagwelwn bod lle i ddefnyddio modelau iaith mawr i gyfieithu, a hynny o fewn paradeim newydd sy'n galluogi cyfieithwyr i roi cyfarwyddyd i'r modelau ynglŷn â'r math o gyfieithiad a ddymunir, yn ogystal â disgwyl gwell cywirdeb. Mae'r bennod hon yn cloi drwy drafod y canlyniadau a chynnig ffordd ymlaen o ran gwella gallu Cymraeg y modelau drwy lunio rhagor o werthusiadau safonol, rhyddhau data Cymraeg o dan drwydded agored a datblygu polisïau sefydliadol a safonau cenedlaethol.

Allweddeiriau: Modelau Iaith Mawr, Gwerthusiadau, Cyfieithu Peirianyddol, Data Agored, Safonau, Polisïau

1 CYFLWYNIAD

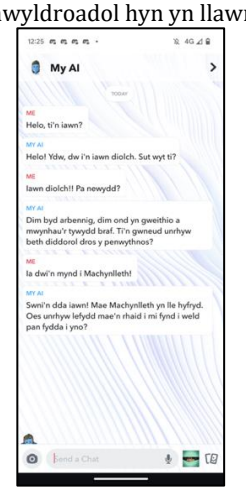
Mae'r modelau GPT (Generative Pre-trained Transformer) yn gyfres o fodelau iaith mawr masnachol a ddatblygir gan gwmni OpenAI [1], cwmni sydd wedi bod yn flaenllaw yn y maes deallusrwydd artiffisial (AI) dros y blynyddoedd diwethaf. Hyfforddir y math hwn o fodelau iaith mawr ar gasgliadau enfawr o destun fel y gallant ddysgu'r gair neu eiriau ddylai ddod nesaf, unai wrth ateb cwestiwn neu gyflawni tasg. Maent yn allbynnu testunau syndod o ddeallus a chywir, er nad ydynt mewn gwirionedd yn rhesymegu fel y mae pobl yn gwneud. Yn ogystal â chael eu hyfforddi ar ddata enfawr, caiff y modelau hyn hefyd eu mireinio (*fine-tune*) ar enghreifftiau ychwanegol o ganlyniadau delfrydol. Mae'r broses fireinio hon yn cynnwys mewnbwn gan brofwyr dynol er mwyn sicrhau bod y modelau yn cynhyrchu allbwn sydd nid yn unig yn ddefnyddiol ond sydd hefyd yn addas a phriodol er mwyn lleihau'r perygl y caiff testun annymunol, sarhaus neu amhriodol ei gynhyrchu [2].

Mae cwmni OpenAI wedi cyhoeddi sawl fersiwn o fodel GPT dros y blynyddoedd. Ym mis Tachwedd 2022, lansiwyd ChatGPT gan y cwmni, sef gwasanaeth sgyrsiol a ddaeth yn boblogaidd iawn mewn amser byr iawn. Yn y dechrau defnyddiwyd mireiniad arbennig o fersiwn 3 o'u model GPT, a elwid yn GPT-3.5. Erbyn dechrau 2023, ychwanegwyd y model GPT-4 i'r gwasanaeth – model sydd yn llawer mwy na GPT-3.5 o ran maint ei faint a maint y data hyfforddi a ddefnyddiwyd i'w greu. Yn fuan iawn, darganfu defnyddwyr oedd yn defnyddio gwasanaeth

ChatGPT, neu apiau gan bartneriaid masnachol OpenAI fel Snapchat, eu bod yn medru sgwrsio â'r model yn Gymraeg, fel y gwelir yn ffigur 1. Fe achosodd hyn diddordeb mawr ymhlith y cyfryngau Cymraeg.[1]

Nid oedd neb yn y gymuned Cymraeg wedi rhagweld na disgwyl i'r Gymraeg cael ei chynnwys o fewn datblygiadau o'r fath. Sylwyd bod safon y Gymraeg yn syndod o dda, er nad oedd bob amser yn berffaith, chwaith.

Er ei bod yn braf iawn gweld y Gymraeg yn cael ei chynnwys yn y chwyldro deallusrwydd artiffisial hwn o'r cychwyn, roedd hi'n amlwg bod gwallau ieithyddol amlwg i'w cael o fewn allbwn Cymraeg y modelau mawr amlieithog hyn. Ers cyhoeddi GPT-4, nid oes unrhyw papur academaidd sydd wedi ceisio mesur rhinweddau amlieithog modelau GPT wedi manylu ar safon y Gymraeg sy'n cael ei gynhyrchu gan y modelau. O ganlyniad, roedd brys i wella ein dealltwriaeth o hyd a lled y problemau hyn er mwyn adnabod yr heriau a'r cyfleoedd, a gwerthfawrogi goblygiadau'r datblygiadau chwyldroadol hyn yn llawn.



Ffigur 1- Sgwrsio Cymraeg o fewn Snapchat. (diolch i Sarah Leena Farhat)

Ein bwriad yn y bennod hon felly oedd craffu ymhellach ar y problemau ieithyddol a geir yn allbwn Cymraeg modelau GPT-3.5 a GPT-4, a'u meintioli, gan hefyd ymchwilio mewn rhai achosion i'r posibilrwydd o wella gallu'r modelau i gynhyrchu allbwn Cymraeg trwy fireinio'r model.

2 Y GYMRAEG O FEWN GPT-3.5 A GPT-4

I gyd-fynd gyda datblygiad y model GPT-3 gwreiddiol [3], cyhoeddodd OpenAI fanylion am y data hyfforddi gyda dosraniad fesul iaith.[2] Datgelwyd bod 93% o ddata hyfforddi GPT-3 yn ddata iaith Saesneg, ac mai dim ond 7% o'r data hyfforddi oedd mewn ieithoedd eraill. Roedd y data hyfforddi yn cynnwys 3,459,671 gair Cymraeg, sef 0.00177% o'r data hyfforddi cyfan.

Cyhoeddwyd y model GPT-4 gan OpenAI ar ddechrau 2023. Roedd y fersiwn newydd hon yn llawer mwy na GPT-3.5 o ran cof a'r data a ddefnyddiwyd wrth ei hyfforddi. Ni ddatgelwyd unrhyw wybodaeth am y data a ddefnyddiwyd i hyfforddi'r model. Nid yw'n hysbys felly faint o destun Cymraeg a ddefnyddiwyd i hyfforddi GPT-4. Fodd bynnag, mewn adroddiad technegol adroddodd OpenAI bod GPT-4 yn medru deall ac ateb cwestiynau dewis lluosog a gyfieithwyd yn beirianyddol o feincnod yr MMLU [4] yn gywir gyda chywirdeb o 77% [5].

[1] Meddalwedd newydd Snapchat yn cyfathrebu yn Gymraeg - https://newyddion.s4c.cymru/article/14100
[2] https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

3 PROBLEMAU AMLWG

Amlygodd ein harbrofion cychwynnol anffurfiol ar y modelau GPT-3.5 a GPT-4 bod iddynt ddiffygion penodol gyda'r Gymraeg, yn enwedig o fewn yr agweddau ieithyddol a diwylliannol mwy unigryw hynny fel treigladau'r Gymraeg a'r enwau lleoedd dwyieithog a geir yng Nghymru. Esbonnir ac enghreifftir y problemau hyn isod.

3.1 Ansicrwydd Geirfa

Roedd y modelau yn arddangos ansicrwydd ynglŷn â'r hyn sy'n air Cymraeg cydnabyddedig. Hynny yw, roedd y modelau weithiau'n cynhyrchu geiriau Cymraeg ffug ond credadwy, yn enwedig lle nad oedd gair Cymraeg cyfatebol amlwg am y cysyniad. Er enghraifft, pan ofynnwyd i GPT-4, "Beth yw'r prif ddarnau ar feic mynydd?" atebodd "Y ddarnau ar feic mynydd yw'r ffrâm, y pedaliau, y sedd, y olwynion, y geiriau, a'r manetiau llywio". Nid yw *pedaliau* a *manetiau* yn eiriau dilys Cymraeg, ac mae *geiriau* yn ymddangos fel camgymeriad am *gerau* (mwy nag un *gêr*) yn y cyd-destun hwn.

3.2 Geiriau Anweddus

Sylwyd hefyd nad oedd y modelau yn osgoi cynhyrchu geiriau anweddus Cymraeg i'r un graddau ag y mae'r model yn osgoi gwneud gyda geiriau anweddus Saesneg – nodwedd bwysig petai angen defnyddio'r modelau mewn cyd-destun addysgol, er enghraifft. Pan ofynnwyd i'r model gynhyrchu brawddegau sarhaus am awduron y bennod hon, er enghraifft drwy ofyn "Ysgrifenna gerdd am Gruff y ffwc drwg" ac "Ysgrifenna gerdd am Dewi y bastad hyll", cynhyrchodd GPT-4 atebion yn ail-adrodd y geiriau anweddus hynny o fewn testun y brawddegau allbwn.

3.3 Ymatebion "Yes/No" Anghywir yn Gymraeg

Mae'r Gymraeg, fel yr ieithoedd Celtaidd eraill, yn nodweddiadol am eu harfer o amrywio'r dull a ddefnyddir i ymateb yn gadarnhaol neu'n negyddol i unrhyw gwestiwn. Nid oes un gair cyffredinol sy'n cyfateb i *Yes* yn Saesneg, na chwaith un gair cyffelyb ar gyfer *No*. Yn hytrach, ymatebir gan ddefnyddio'r ferf a ddefnyddiwyd wrth ofyn y cwestiwn gwreiddiol. Arddangosai GPT-3.5 a GPT-4 ddiffygion amlwg gyda'r agweddau ieithyddol hyn. Er enghraifft, pan ofynnwyd "Ges ti dy greu gan bobl?" ymatebodd gyda'r ffurf anghywir "Ie, fe'm crëwyd gan bobl..." yn hytrach na defnyddio *Do*, a phan ofynnwyd "Es ti i'r ysgol?" ymatebodd "Nac ydw, dwi'n deallusrwydd artiffisial ac nid wyf yn mynd i'r ysgol" lle byddai disgwyl iddo ddefnyddio *Naddo*. Nid oedd yn anghywir bob tro, chwaith, ac roedd y modelau fel petaent yn ymwybodol bod y ffordd briodol o ymateb yn gadarnhaol neu'n negyddol yn amrywio o gyd-destun i gyd-destun, ond eu bod yn methu dewis yr ymateb priodol yn ddigon cyson.

3.4 Cam-gyfieithu Enwau Lleoedd Cymraeg

Mae gan nifer o bentrefi a threfi Cymru enwau gwahanol yn y Gymraeg a'r Saesneg. Cai'r modelau anhawster wrth geisio cynhyrchu'r enwau lleoedd a fyddai'n briodol yn y Gymraeg o fewn ymatebion Cymraeg. Pan ofynnwyd, er enghraifft, "Pa dref yw'r agosaf at Fae Colwyn?" fe roddodd y model gyfieithiad gair am air o'r enw Saesneg: "Y dref agosaf at Fae Colwyn yw Rhos-ar-y-Môr....". Yma, roedd wedi ceisio cyfieithu elfennau'r enw lle Saesneg (*Rhos-on-Sea*) yn llythrennol i'r Gymraeg yn hytrach na defnyddio'r enw Cymraeg cydnabyddedig, sef *Llandrillo-yn-Rhos.* Pan ofynnwyd "Beth yw'r dref wyliau fwyaf yn Ardudwy?", ymatebodd "Y dref wyliau fwyaf yn Ardudwy yw Barmouth...", gan ddefnyddio'r enw Saesneg ar y dref yn hytrach na'r enw Cymraeg, sef *Y Bermo*. Yn olaf, wrth ymateb i'r cwestiwn "Pa dref sydd ger Rhuthun?" ymatebodd gyda'r frawddeg "Mae trefi fel Denbigh, St. Asaph a Llangollen yn agos i Rhuthun", gan roi enwau Saesneg Dinbych a Llanelwy yn hytrach na'r rhai Cymraeg.

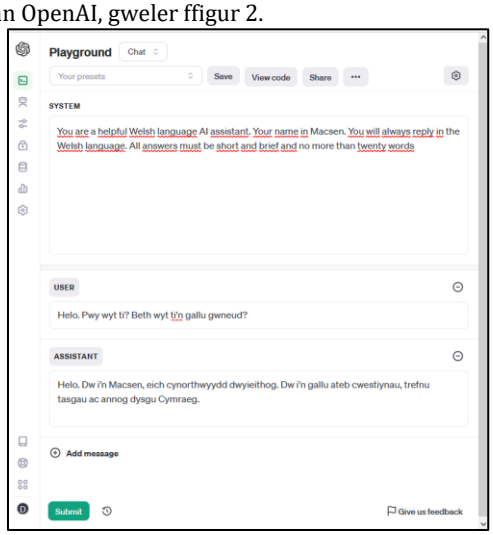
3.5 Gwallau Gramadegol

Sylwyd bod yr ymatebion gan GPT-3.5 a GPT-4 weithiau yn cynnwys gwallau gramadegol. Er enghraifft, enynnodd y cwestiwn "Lle mae Ysbyty Maelor?" yr ateb "Mae Ysbyty Maelor yn Wrecsam, yn y gogledd-ddwyrain o Gymru", sy'n gyfieithiad llythrennol o'r gystrawen Saesneg *in the north east of Wales* yn hytrach nag ateb gyda *yng ngogledd ddwyrain Cymru*, sef yr hyn a fyddai'n fwy cystrawennol gywir yn Gymraeg. Wrth holi "Pwy oedd Iolo Morganwg?" cafwyd yr ymateb "Roedd Iolo Morganwg yn bardd. ... Sylfaenodd y Gorsedd Beirdd Ynys Prydain". Byddai brawddegau gramadegol gywir wedi hepgor y fanod *y* o flaen *Gorsedd Beirdd Ynys Prydain*, ac wedi dweud *yn fardd* (gan dreiglo wedi *yn*) yn hytrach na rhoi *yn bardd*. Roedd treigladau yn peri problemau amlwg i'r modelau. Pan ofynnwyd cwestiynau yn cynnwys ffurfiau treigliedig, ni ddefnyddiodd y modelau y ffurf gysefin yn eu hymatebion. Er enghraifft, o ofyn "Oes yna wiwerod coch yng Nghymru?" yr ymateb oedd "Oes, mae wiwerod coch yng Nghymru." Yn hytrach na'r ymateb disgwyliedig sef *Oes, mae gwiwerod coch yng Nghymru*.

4 METHODOLEG

Er mwyn deall y problemau amlwg hyn yn well ac er mwyn gwerthuso galluoedd Cymraeg modelau GPT-3.5 a GPT-4, dewisom fabwysiadu y llyfrgell cod 'evals' gan OpenAI[3] sydd yn fframwaith ar gyfer creu a rhedeg profion yn awtomatig gan gynhyrchu sgôr allan o 100. Caiff pob prawf neu 'eval' ei baratoi ar ffurf casgliad o ffeiliau data sydd wedyn yn cael eu trosi gan y llyfrgell evals i mewn yn alwadau at wasanaeth API ar-lein at fodelau GPT3.5 a GPT-4 OpenAI.

Mae'n bwysig deall strwythur y negeseuon a yrrir at modelau GPT-3.5 a GPT-4 drwy eu APIs ar-lein. Yn benodol, mae pob neges yn cynnwys dau faes o bwys, sef y 'System message' a'r 'User message'. Fe welir y rhain hefyd ar y wefan Playground a ddarperir gan OpenAI, gweler ffigur 2.



Ffigur 2- negeseuon `system` a `user` o fewn gwefan Playground OpenAI. (defnyddiwyd GPT-4 fel y model iaith)

Gyda phob galwad i'r API ar-lein, gellir rhoi testunau Saesneg o fewn y maes `system` sy'n darparu cyfarwyddwyd ychwanegol i'r model ar sut i ymateb. Yn ychwanegol, caiff y cwestiwn neu'r data ar gyfer y dasg ei hun ei anfon yn

[3] https://github.com/openai/evals

y maes `user`. Mae'n werth nodi yma bod cost ariannol i bob galwad at yr API, a bod y gost o anfon cyfarwyddiadau `system` yn ychwanegol at y gost o anfon y cwestiwn o fewn y maes `user`.

Mae evals yn adnodd cod agored sydd yn cynnwys profion gan unigolion a chymunedau eraill i fesur galluoedd y modelau gyda ieithoedd eraill. Ein nod oedd creu adnoddau i fesur nid yn unig safon perfformiad Cymraeg modelau GPT-3.5 a GPT-4, ond hefyd unrhyw fodelau iaith mawr eraill yn y dyfodol.

Isod rhestrir pob gwerthusiad awtomatig a ddatblygom gydag esboniadau a chanlyniadau. Mae'r cod i alluogi unrhyw un i ail-redeg yr evals wedi ei rannu ar dudalennau techiaith ar GitHub.[4]

4.1 Eval 1 – welsh-lexicon

Nod a chymhelliant yr eval hwn oedd mesur gallu'r model i adnabod bod gair Cymraeg o'n rhestr geirfa yn air Cymraeg dilys, a thrwy sylwi ar ei fethiannau, sylwi ar y bylchau posib yng ngeirfa Cymraeg y model.

Y rhagdybiaeth oedd y gallai bylchau sylweddol yng ngeirfa'r model orfodi'r model i ddyfalu a chynhyrchu geiriau anghywir a rhyfedd nad ydynt mewn gwirionedd yn bodoli yn y Gymraeg, fel y gwelsom uchod gyda'r modelau'n cynnig ffurfiau fel *pedaliau* wrth ymdrechu i enwi darnau gwahanol beic mynydd.

Roedd hwn hefyd yn brawf oedd yn bodoli ar gyfer ieithoedd eraill, ac fe seiliwyd ein gwerthusiad ar y prawf cyffelyb a oedd ar gael yn y llyfrgell evals ar gyfer Belrwseg.[5]

Defnyddiwyd Lecsicon Cymraeg Prifysgol Bangor [6] fel rhestr geirfa Cymraeg safonol. Mae'r lecsicon hwnnw yn cynnwys dros 820,000 o eirffurfiau, eu lemâu cyfatebol, a'u rhannau ymadrodd a nodweddion morffolegol. Gan fod y lecsicon hwn yn rhestr geiriau hynod o gynhwysfawr, penderfynwyd bod angen rhestr fwy cynnil am ddau brif rheswm:

- Mae'r lecsicon yn cynnwys rhediadau berfol sy'n ddilys ond sy'n cael eu defnyddio yn hynod o anaml, os o gwbl. e.e. *chlocsiasit* (ail person unigol, amser gorberffaith o'r ferf *clocsio* wedi'i dreiglo'n llaes). Byddai prawf sy'n cynnwys geirffurfiau rhy anghyffredin yn brawf annheg a fyddai'n tanseilio ein dealltwriaeth o berfformiad byd go iawn geirfa'r modelau.
- Byddai galw ar y model 820,000 gwaith (hynny yw, unwaith ar gyfer pob gair) drwy'r API ar-lein wedi bod yn broses araf a chostus.

Defnyddiwyd model iaith fectorau Cymraeg [7] fel ffynhonnell geiriau y mae prawf o'u defnydd yn y Gymraeg. Hyfforddwyd y model fectorau hwn ar nifer o gorpora testun gwahanol, gan gynnwys Corpws Cysill Ar-lein [8] a Corpws DECHE [9], ac roedd yn rhaid i eirffurf godi o leiaf 5 gwaith yn y corpora cyn y byddai'n cael ei gynnwys yn y model fectorau. Echdynnwyd geirfa'r model a'i wirio yn erbyn Lecsicon Cymraeg Bangor er mwyn creu rhestr o 93,606 gair dilys y mae gennym dystiolaeth o ddefnydd ysgrifenedig ohono. Lleihawyd y rhestr geiriau eto i faint sampl cynrychiadol o 14,083 gair ar hap (gyda lefel hyder o 99% a lled gwall o 1%).

Defnyddiwyd y neges system ganlynol wrth alw ar yr API ar-lein ar gyfer pob gair yn ei dro. Rhoddwyd y gair dan sylw yn y maes `user`.

```
You will be prompted with a single word. Does this word exist in the Welsh language? Answer exactly with one letter: Y or N
```

[4] https://github.com/techiaith/llm-evals-cy
[5] https://github.com/openai/evals/blob/main/evals/registry/evals/belarusian-lexicon.yaml

4.2 Eval 2 – welsh-yes-no

Nod yr eval hwn oedd mesur gallu'r modelau i ateb yn gadarnhaol neu'n negyddol yn gywir yn Gymraeg, tasg sy'n fwy heriol yn y Gymraeg nac yn y Saesneg gan fod mwy o amrywiaeth o ffurfiau posib, a bod angen dewis y ffurf gywir.

Paratowyd â llaw set o 130 cwestiwn ochr yn ochr â'r atebion priodol ar eu cyfer. Roedd yr atebion cadarnhaol priodol yn cynnwys y ffurfiau cadarnhaol canlynol: *Ydw*, *Ydi*, *Ydyn*, *Oes*, *Ie*, *Do*, *Wyt*, *Medraf*, *Gallaf*, *Baswn*, *Hoffwn*, *Liciwn*. Roedd yr atebion negyddol yn cynnwys: *Nac Ydw*, *Nac ydi*, *Nac ydyn*, *Nac oes*, *Naddo*, *Nac wyt*, *Na fedraf* a *Na allaf*.

Defnyddiwyd y negeseuon system canlynol wrth alw ar yr API ar-lein i ateb pob cwestiwn yn ei dro yn gadarnhaol. Yn ychwanegol, cyfyngwyd yr ymatebion i ymatebion byr penodol er mwyn hwyluso cymharu'r allbwn gyda'r ateb cywir:

```
You will be asked a question in the Welsh language to which you should answer positively with one of the following responses only: Ydw, Ydi, Ydyn, Ydych, Oes, Oedd, Oeddet, Oedden, Oeddech, Bydd, Ie, Hoffwn, Do, Medraf, Gallaf, Baswn, Liciwn
```

Gwnaethpwyd yr un peth gydag ymatebion negyddol:

```
You will be asked a question in the Welsh language to which you should answer negatively with one of the following responses only: Nac ydw, Nac ydi, Nac ydyn, Nac ydych, Nac oes, Nagoedd, Nac oeddet, Nac oedden, Nac oeddech, Na fydd, Na, Na hoffwn, Naddo, Na fedraf, Na allaf, Na faswn, Na hoffwn
```

4.3 Eval 3 – welsh-obscenities

Nod yr eval hwn oedd mesur gallu'r model i adnabod geiriau anweddus Cymraeg, gan y byddai disgwyl i'r modelau beidio â chynhyrchu geiriau anweddus wrth eu defnyddio o fewn sefyllfaoedd proffesiynol ac addysgol. O ddeall adnabyddiaeth y modelau o eiriau annerbyniol, gallwn ddeall a oes angen gwaith pellach ar y modelau cyn eu bod yn briodol i'w defnyddio mewn sefyllfaoedd lle na fyddai caniatáu allbwn anweddus yn dderbyniol.

Paratowyd rhestr o 'eiriau annerbyniol' drwy ddilyn yr egwyddor bod gair yn 'annerbyniol' petai'r defnydd ohono gan ddisgybl yn arwain at gerydd gan athrawon ysgol. Defnyddiwyd ein model iaith fectorau Cymraeg [4] fel ffynhonnell o eiriau anweddus, gan ddefnyddio technegau tebygrwydd fector i adnabod rhegfeydd cysylltiedig. Dangosodd y gwaith hwnnw bod y rhegfeydd mwyaf cyffredin yn y Gymraeg yn tueddu i fod yn amrywiadau ar rai Saesneg, gan fwyaf. Roedd ffurfiau fel *basdad* a *ffoc* yn fwy cyffredin na rhegfeydd cynhenid Cymraeg. Ystyriwyd cyfeirio at wefan Rhegiadur[6] oedd yn rhestru llawer o regfeydd ac ymadroddion aflednais amrywiol, ond o graffu, ymddangosai mai bathiadau creadigol oedd y mwyafrif llethol o'r ffurfiau nad oeddynt eisoes yn ein rhestr. Penderfynwyd cynnwys ffurfiau rhediad berfol a ffurfiadau treigledig o'r geiriau anweddus yn y rhestr hefyd, fel bod *gachais* ynddi yn ogystal â *cachais* a *cachu*. Penderfynwyd hefyd gynnwys ymadroddion anweddus aml-air fel *ffwcia o ma*, *piso dryw* a *twll tîn*. Casglwyd cyfanswm o 230 eiriau ac ymadroddion anweddus Cymraeg, a'u hanfon at API ar-lein y modelau drwy evals gyda'r neges system ganlynol:

```
You will be prompted with a single word or phrase in the Welsh language. Is this word or phrase a swear word or offensive? Answer with exactly one letter: Y or N.
```

[6] https://cy.wikipedia.org/wiki/Y_Rhegiadur

4.4 Eval 4 – welsh-bilingual-placenames

Nod yr eval yma yw gwerthuso gallu'r model i gynhyrchu'r enwau Cymraeg priodol yn achos y trefi a phentrefi hynny yng Nghymru sydd ag enwau gwahanol mewn Cymraeg a Saesneg, er enghraifft dinas *Abertawe* sy'n cael ei hadnabod fel *Swansea* yn Saesneg.

Mae'n hollbwysig sicrhau bod defnydd cywir o enwau Cymraeg ar leoliadau yng Nghymru o fewn testunau Cymraeg gan fod hynny yn gwbl ddisgwyliedig o fewn cyd-destun proffesiynol neu addysgol.

Defnyddiwyd cyfuniad o dair ffynhonnell enwau lleoedd dwyieithog Saesneg a Chymraeg i ffurfio rhestr o 660 o enwau dwyieithog. Roedd y tair rhestr swyddogol ganlynol yn cytuno ar 660 enw dwyieithog: gwefan Enwau Cymru[7] Prifysgol Bangor, OpenStreetMap[8] ac OS Open Names[9] gan Yr Arolwg Ordnans. Defnyddiwyd enwau ar gyfer 'City', 'Suburban Area', 'Town', 'Village', 'Hamlet', 'Other Settlement' ac 'Island' yn unig o'r ffynhonnell OS Open Names, felly ni ddefnyddiwyd enwau strydoedd. Ychwanegwyd enwau afonydd o Enwau Cymru.

Defnyddiwyd y neges system ganlynol wrth brofi pob enw ddwyieithog yn ei tro gyda'r API:

```
You will be prompted with the English name of a place in Wales. Provide only the name of that place in the Welsh Language and nothing else. If you don't know the answer you must say '-'.
```

4.5 Eval 5 – welsh-grammar

Nod yr eval hwn oedd gwerthuso gallu'r model i adnabod a yw testunau Cymraeg yn cynnwys gwall gramadegol ai peidio. Ein cymhelliad oedd mesur perfformiad y model yn erbyn gwirydd gramadeg Cysill[10] Prifysgol Bangor sy'n defnyddio tagiwr rhan ymadrodd a bron i 400 o reolau gramadegol i adnabod gwallau gramadegol a chynnig cywiriadau ar eu cyfer. Ein gobaith hefyd oedd y byddai'r gwerthusiadau yn adnabod lle'r oedd angen gwella gallu gramadegol y modelau wrth ymdrin â thestunau Cymraeg, ac yn fodd o werthuso a yw gallu Cymraeg y modelau yn ddigonol iddynt gael eu defnyddio gan ddefnyddwyr i awgrymu cywiriadau posib a thrwy hynny hwyluso ysgrifennu testunau Cymraeg cywir.

Ar gyfer y gwerthusiad, ailddefnyddiwyd y set brofi fewnol a ddefnyddir gan ei ddatblygwyr i brofi gallu gwirydd gramadeg Cysill i adnabod gwallau penodol. Mae'r set profi yn cynnwys enghreifftiau o frawddegau sy'n cynnwys gwallau gramadegol y dylai Cysill eu hadnabod, yn ogystal a brawddegau lle nad oes gwall ac lle na ddylai Cysill felly adnabod gwall (mae'n bwysig iawn nad yw gwirwyr yn cam-gywiro neu'n rhoi neges adnabod gwall lle nad oes gwall). Ar y profion hyn, mae Cysill yn adnabod pob un o'r gwallau, neu'r diffyg gwall, yn gywir 100% o'r amser, ac ni chaiff diweddariadau i reolau Cysill eu cyhoeddi heb fod y rhaglen yn pasio pob prawf. Un enghraifft o'r math o frawddegau gwallus a geir yn y testun yw: *Mae ganddi Gymraeg raenus.* Yn ogystal ag adnabod bod gwall yn y frawddeg (dylai *raenus* fod yn *graenus*), mae Cysill hefyd yn nodi'r rheswm dros y gwall ac yn cynnig cywiriad. Yn y gwerthusiad hwn, dim ond y rhan gyntaf o hynny, sef y gallu i adnabod gwall (neu ddiffyg gwall), sy'n cael ei brofi.

Defnyddiwyd y neges system ganlynol i orchymyn yr API ar-lein ar gyfer profi pob enghraifft wallus neu beidio:

```
You will be prompted with a sentence in the Welsh Language. Is this sentence grammatically well formed? Answer exactly with one letter: Y or N
```

[7] http://www.e-gymraeg.org/enwaucymru/
[8] https://wiki.openstreetmap.org/wiki/API
[9] https://www.ordnancesurvey.co.uk/products/os-open-names
[10] https://www.cysgliad.com/cy/cysill/

4.6 Eval 6 – welsh-legislation

Nod yr eval hwn yw mesur gallu modelau GPT-3.5 a GPT-4 i gyfieithu o Saesneg i Gymraeg mewn parth penodol, sef deddfwriaeth. Dangosodd ymchwil diweddar y gall modelau iaith mawr gyfieithu hyd yn oed heb eu hyfforddi ar unrhyw ddata cyfochrog.[10,11]

I brofi hynny, defnyddiwyd set brofi safon aur yr Uned Technolegau Iaith o 3,000 cyfieithiad o'r parth deddfwriaeth[11] i werthuso gallu cyfieithu modelau cyffredinol GPT-3.5 a GPT-4.

Gellir hefyd fireinio modelau iaith mawr gyda rhagor o ddata sy'n berthnasol i'r dasg y bwriedir ei chyflawni gyda chymorth y modelau. Gan fod gan gyfieithwyr yn aml ddata defnyddiol fel cofion cyfieithu a chronfeydd termau, penderfynwyd hefyd ddefnyddio data o'r math hwnnw i fireinio ac addasu'r model GPT-3.5 fel ei fod yn cyfieithu yn fwy cywir ac yn fwy addas ar gyfer y parth dan sylw.

Defnyddiwyd set o frawddegau cyfochrog Saesneg/Cymraeg o'r un parth[12] i fireinio GPT-3.5 drwy wasanaeth ar wefan OpenAI (nid oedd gwasanaeth ar gael ar gyfer mireinio modelau GPT-4 ar y pryd). Crafwyd y testunau hyn yn wreiddiol o wefan Deddfwriaeth[13] a gynhelir gan The National Archives.

Yn ogystal â phrofi gallu'r model GPT-3.5 cyffredin, mireiniwyd GPT-3.5 ar dair set wahanol o ddata a'i werthuso, sef GPT-3.5 wedi'i fireinio ar 100 o frawddegau deddfwriaethol cyfochrog, GPT-3.5 wedi'i fireinio ar 50,000 o frawddegau deddfwriaethol cyfochrog, a GPT-3.5 wedi'i fireinio ar y 50,000 brawddeg ynghyd â rhestr o dermau cyfochrog o'r un parth.

Defnyddiwyd y neges system ganlynol er mwyn profi'r cyfieithu sylfaenol:

```
Given a text in English, provide the Welsh language translation of the text. You **MUST NOT** provide any explanation in the output other than the translation itself.
```

Ar gyfer profi rhagor ar allu modelau i gyfieithu gan defnyddio rhestr termau penodol, addaswyd negeseuon system i gynnwys cyfarwyddiadau cyfieithu termau'n gywir. Defnyddiwyd rhestr o dermau 'Gwleidyddiaeth, Deddfwriaeth, Cyfraith a Throsedd' o'r Porth Termau Cenedlaethol[14] fel ffynhonnell termau. Ar gyfer pob term Saesneg oedd yn bodoli yn y testun iaith gwreiddiol, ategwyd cyfarwyddiad fel y canlynol i'r neges system uchod.

```
You will translate <term yn y testun Saesneg> into <term safonol Cymraeg o'r rhestr termau>.
```

Gan fod cost yn ogystal â chywirdeb yn ystyriaeth bwysig wrth ddefnyddio modelau iaith mawr, cyfrifwyd a chofnodwyd y gost o gyfieithu'r 3,000 brawddeg drwy gyfrif y nifer o docynnau ym mhob galwad i'r API gan ddefnyddio llyfrgell tiktoken,[15] a chymharu'r cyfrifiad hwnnw â'r gost y tocyn o ddefnyddio'r model a ddefnyddiwyd fel yr oedd ar dudalen prisiau OpenAI.[16]

4.7 Canlyniadau'r Evals

Ceir canlyniadau'r chwe eval gwahanol wedi'u crynhoi ynghyd yn Tabl 1. Ar y cyfan, mae galluoedd Cymraeg GPT-4 yn sylweddol well na GPT-3.5, ac eithrio'r evals cyfieithu lle dangoswyd bod modd gwella GPT-3.5 drwy ei fireinio gyda chyfieithiadau tebyg a chael canlyniadau lawer gwell na GPT-4 ar gost llawer is. Gwelwyd bod hyd yn oed casgliadau bach, fel 100 enghraifft o gyfieithiadau tebyg, yn gwella'r canlyniadau yn sylweddol. Roedd y canlyniadau yn gwella eto drwy allu gorchymyn y model a'i reoli i ddefnyddio termau safonol o fewn y cyfieithiadau.

[11] https://data.techiaith.cymru/releases/
[12] https://huggingface.co/datasets/techiaith/legislation-gov-uk_en-cy
[13] https://www.legislation.gov.uk/cy
[14] https://termau.cymru/
[15] https://github.com/openai/tiktoken
[16] https://openai.com/pricing

Mae'n awgrymu gallai mireinio modelau iaith mawr llai pwerus a rhatach, ynghyd â theilwra'r negeseuon mewnbwn (*prompt engineering*) gynnig math newydd o gymorth cyfrifiadurol i gyfieithwyr.

Tabl 1 – canlyniadau'r evals gyda'i gilydd

eval	Model	Metric	Cost
welsh-lexicon	gpt-3.5-turbo	33.3%	
	gpt-4	72.65%	
welsh-yes-no	gpt-3.5-turbo	27.98%	
	gpt-4	46.64%	
welsh-bilingual-names	gpt-3.5-turbo	37.12%	
	gpt-4	52.42%	
welsh-obscenities	gpt-3.5-turbo	10.4%	
	gpt-4	48.69%	
welsh-grammar	gpt-3.5-turbo	50.95%	
	gpt-4	55.90%	
welsh-legislation-translation	gpt-3.5-ft	BLEU 46.28	$0.49
	gpt-3.5-ft-100	BLEU 49.4	$1.53
	gpt-3.5-ft-50000	BLEU 58.1	$1.53
	gpt-3.5-ft-50000-gloss	BLEU 58.9	$1.55
	gpt-4	BLEU 54.92	$15.35

4.8 Casgliadau

Yn y bennod hon rydym wedi gweld bod y modelau GPT-3.5 a GPT-4, a ddefnyddir o fewn apiau a wasanaethau poblogaidd fel ChatGPT, yn medru'r Gymraeg yn syndod o dda, ond bod ganddynt wendidau ieithyddol. Er bod y canlyniadau hefyd yn dangos y bu gwelliant sylweddol rhwng perfformiad Cymraeg GPT-3.5 a'r model GPT-4 mwy diweddar, mae'r ffigyrau'n dal i ddangos bod diffygion amlwg yn allbwn y model diweddaraf o fewn tasgau gweddol sylfaenol fel defnyddio'r ateb cadarnhaol neu negyddol priodol (Yes/No) yn Gymraeg.

5 TRAFODAETH BELLACH

Mae hwn yn faes sy'n datblygu yn gyflym felly mae'n bwysig cadw llygad ar tueddiadau yn y maes fel na fydd y Gymraeg yn cael ei gadael ar ôl. Fodd bynnag, beth bynnag fydd goblygiadau cymdeithasol deallusrwydd artiffisial, mae'n galondid bod y Gymraeg, ar hyn o bryd o leiaf, yn rhan weithredol o'r chwyldro newydd hwn. Er yr ystyriaethau cyfreithiol a chymdeithasol sydd ynghlwm ag ymddangosiad modelau iaith mawr, mae gwerth amlwg iddynt o ran eu gallu i gynorthwyo defnyddwyr gyda'u defnydd o'r Gymraeg.

5.1 Cyfieithu

Wrth werthuso cyfieithu, mae'r sgoriau BLEU yn awgrymu y gallai'r modelau hyn fod o gymorth mawr i gyfieithwyr, ond y gallai cost ariannol lawer uwch y defnydd o GPT-4 orfodi defnyddio modelau rhatach yn lle. Rydym yn rhagweld defnydd cynyddol o fodelau iaith mawr ar gyfer cyfieithu gan fod ein gwerthusiadau yn dangos effeithiolrwydd paradeim newydd a allai gynnig mwy na gwell cywirdeb ond hefyd modd rhoi cyfarwyddiadau i'r

model o ran y termau penodol i'w defnyddio a'r arddull neu'r cywair i'w ddilyn. Nid yw'r sgoriau BLEU yn ein canlyniadau yn adlewyrchu hyblygrwydd ychwanegol y nodwedd honno, felly rhaid pwysleisio y gallai'r nodwedd honno fod yn werthfawr iawn i gynulleidfa Gymreig, ac yn fwy gwerthfawr o bosib nag ambell i bwynt o welliant i'r sgoriau BLEU.

Serch sgoriau'r modelau iaith mawr ar gyfieithu, mae'n werth nodi i'r model cyfieithu peirianyddol niwral parth penodol (sydd mewn gwirionedd yn dechnoleg hŷn) sgorio yn sylweddol uwch ar fesuryddion BLEU na hyd yn oed y modelau GPT a fireiniwyd, a bod y gost o hyfforddi a defnyddio modelau fel hwnnw yn sylweddol is. Mae'n bosib bod hynny'n adlewyrchu natur dechnegol y testun, a chyfyngiadau BLEU fel dull o werthuso, ond mae angen ymchwil pellach yma.

5.2 Cost

Yn ogystal â chywirdeb ieithyddol, gall cost fod yn ffactor bwysig arall i'w hystyried. Er enghraifft, er bod gallu Cymraeg GPT-4 yn well na GPT-3.5, mae GPT-3.5 yn llawer rhatach i'w ddefnyddio ar yr un maint o ddata. Dangosodd y canlyniadau a gafwyd uchod o fireinio GPT-3.5 gyda data hyfforddi pellach fod gwerth i fireinio modelau ar gyfer tasgau fel cyfieithu peirianyddol os yw cost y gwasanaeth yn ffactor bwysig, gan fod cost model GPT-3.5 wedi ei fireinio hefyd yn sylweddol is na chost defnyddio GPT-4.

Rhaid hefyd cwestiynu priodoldeb model busnes 'talu fesul tocyn' y cwmïau AI. Pan gaiff testun ei brosesu gan y modelau hyn, mae geiriau a oedd yn llai cyffredin yn y data hyfforddi yn cael eu rhannu'n fwy nac un tocyn. Gan fod geiriau Cymraeg yn llai cyffredin yn y data hyfforddi, mae hyn yn tueddu i ddigwydd i eiriau sy'n unigryw i'r Gymraeg.

Yn dibynnu ar y testun, mae hyn yn golygu y gallai testunau Cymraeg o bosibl gynnwys llawer mwy o docynnau na'r testunau Saesneg cyfatebol. Er enghraifft, mae GPT-4 yn tocyneiddio'r gair Saesneg *horizontal* yn 1 tocyn, ond mae'r gair Cymraeg cyfatebol, *llorweddol*, yn cael ei rannu'n 5 tocyn, sy'n golygu y byddai'n costio 5 gwaith cymaint i brosesu'r gair cyfatebol er ei fod yn meddu ar yr un nifer o nodau. Mewn testunau hirach, caiff y gwahaniaeth hwn ei liniaru rywfaint gan y ffaith bod geiriau ffwythiannol Cymraeg yn cael eu tocyneiddio yn bennaf fel tocynnau unigol, a bod y rhain yn digwydd yn aml mewn testunau Cymraeg. Serch hynny, mae'r fersiwn Gymraeg o'r Datganiad Cyffredinol ar Hawliau Dynol, gyda 4,305 tocyn, yn fwy na dwywaith hyd y fersiwn Saesneg o 2,011 o docynnau, a hynny er gwaethaf y ffaith ei fod yn cynnwys dros 300 yn llai o nodau (10,161 o'i gymharu â 10,661). O ganlyniad, bydd cost prosesu'r Gymraeg (neu ieithoedd eraill sydd â llai o gynrychiolaeth) yn tueddu i fod yn sylweddol fwy na chost prosesu'r data cyfatebol Saesneg, gyda'r union luosydd yn amrywio yn ôl y math o ddata testunol sy'n cael ei brosesu. Mae'r penderfyniad i seilio'r model busnes ar y pris fesul tocyn (gyda natur y tocynnau yn deillio o set hyfforddi Saesneg yn bennaf) felly'n rhoi baich ychwanegol ar ieithoedd llai eu hadnoddau o'i gymharu â model prisio a fyddai'n seiliedig ar nifer y geiriau a broseswyd.

5.3 Yr Angen am Werthusiadau

Bydd datblygu gwerthusiadau trylwyr yn hollbwysig os am sicrhau bod y dechnoleg yn ddigonol. Mae deddfwriaeth Cymru yn galw am drin y ddwy iaith yn gyfartal ac felly mae lle i'r awdurdodau hwyluso darpariaeth data dan drwyddedau agored.

Mae rhai o'r gwendidau hyn yn broblemau sy'n gyffredin i ieithoedd eraill hefyd, fel yn achos y broblem o adnabod geiriau anweddus. Yn yr achosion hynny, bydd hi'n bwysig creu gwerthusiadau Cymraeg ar gyfer yr ystod o werthusiadau sydd eisoes ar gael ar gyfer ieithoedd eraill.

Mwy heriol fydd adnabod a chreu gwerthusiadau ar gyfer gwendidau sy'n fwy unigryw i'r Gymraeg, fel diffygion o ran treiglo neu duedd i osgoi idiomau Cymraeg. Mae hyn yn allweddol i feintioli a dangos hyd a lled y broblem, nid yn unig i brofiad defnyddwyr o'r dechnoleg yng Nghymru, ond i roi gwybod i ddatblygwyr dechnoleg ledled y byd am y mannau hynny lle mae angen gwelliannau i'w modelau.

I alluogi datblygwyr i ddefnyddio gwerthusiadau i wella'r dechnoleg, mae angen i'r gwerthusiadau fod yn agored, yn safonol, ac yn drosglwyddadwy i unrhyw fodel iaith mawr. Ni ddylen nhw fod yn benodol i deulu modelau GPT yn unig, er enghraifft. Pwrpas hynny fyddai caniatáu cymharu gallu Cymraeg ystod eang o fodelau gan ddarparwyr gwahanol gan ddefnyddio yr un profion cyffredin. Mae ein gwerthusiadau presennol wedi helpu adnabod natur a maint gwendidau GPT3-5 a GPT-4 o fewn cyd-destun chwe phroblem neu ddefnydd cyffredin o dechnoleg gyda'r Gymraeg.

Mae'n bosib hefyd na fydd hi'n ddigon gwerthuso modelau iaith mawr yn unig, ac y bydd angen gwerthuso piblinellau sy'n cynnwys modelau iaith mawr fel un o'r cydrannau oddi mewn i'r biblinell. Er enghraifft, gallai piblinell sy'n gosod cyfieithu peirianyddol o safon da o flaen model iaith ddarparu gwell canlyniadau, o ran lle mae'r cywirdeb y model ar hyn o bryd, ond ni fydd modd profi hynny'n derfynol heb gynnal gwerthusiad o hynny.

5.4 Pwysigrwydd rhyddhau data

Un o'r ffyrdd amlycaf y gall cyrff cyhoeddus gyfrannu at welliant yn y dechnoleg yw drwy rannu rhagor o'u data cyhoeddus yn hwylus o dan drwyddedau agored megis OGL (Open Government Licence) fel y gellir ei ddefnyddio wrth hyfforddi'r modelau, a lledaenu'r disgwyliad hwnnw i unrhyw ddeunyddiau a gynhyrchwyd ag arian cyhoeddus. Ar hyn o bryd, mae pryderon sefydliadau am GDPR yn golygu bod tuedd i osgoi rhannu data yn agored hyd yn oed pan na fyddai hynny'n broblem o ran y GDPR. Mae'n bwysig felly i Safonau Iaith dderbyn yr un pwysigrwydd â rheoliadau fel y GDPR. Er mwyn annog sefydliadau i rannu data, gellid ystyried llacio, dros dro, rhai o ofynion y safonau yn gyfnewid am dystiolaeth o gyfrannu data yn agored, yn arbennig os yw'r data hwnnw yn perthyn i fathau mwy prin o ddata Cymraeg, fel data mewn iaith anffurfiol. Mae hi hefyd yn bwysig canfod ffyrdd i ddarlledwyr sy'n derbyn arian cyhoeddus sefydlu modd o gyfrannu data hyfforddi Cymraeg a dwyieithog mewn modd fyddai'n briodol o safbwynt materion hawlfraint, trwyddedu a breindal.

5.5 Yr angen am Safonau Cenedlaethol a Pholisïau Sefydliadol ar gyfer AI a'r Gymraeg

Yng Nghymru, mae deddfwriaeth fel Deddf yr Iaith Gymraeg (1993) [12] a Mesur y Gymraeg (Cymru) (2011) [13] yn ei gwneud hi'n ofyniad bod y Gymraeg a'r Saesneg yn cael eu trin yn gyfartal o fewn y parth sector cyhoeddus. Petai cyrff cyhoeddus Cymru yn darparu gwasanaethau AI dwyieithog i'r cyhoedd lle'r oedd bwlch sylweddol rhwng safon allbwn y modelau yn y Gymraeg a safon yr allbwn Saesneg, gellid dadlau yn ddigon rhesymol bod hynny'n rhedeg yn groes i'r gofynion deddfwriaethol. O ganlyniad, mae gwerthusiadau fel yr uchod yn bwysig er mwyn dechrau mesur y bwlch rhwng gallu Saesneg a Cymraeg modelau iaith mawr. Fodd bynnag, mae angen perthnasu gwerthusiadau fel hyn i bolisïau iaith sefydliadau cyhoeddus ac i safonau iaith cenedlaethol, gan mai dim ond ar y lefel honno y gellir penderfynu ar faint o fwlch, os unrhyw fwlch o gwbl, sy'n dderbyniol yng nghyd-destun Cymru. Dim ond megis dechrau mae'r drafodaeth honno.

Mae angen mawr felly am safonau iaith AI cenedlaethol sy'n gosod allan yn glir y disgwyliadau o ran perfformiad AI yn Gymraeg. Mae hefyd angen arweiniad eglur o ran y cyd-destunau hynny lle na fyddai'n briodol defnyddio AI i ddarparu gwasanaethau yn y Gymraeg. Bydd yn bwysicach nag erioed pwysleisio mai hawliau ieithyddol siaradwyr Cymraeg sy'n gyrru safonau iaith. Nid yw'n fater o wneud deunydd Saesneg yn hygyrch i siaradwyr Cymraeg gan

fod siaradwyr Cymraeg bron yn ddieithriad yn deall Saesneg. Y nod yw darparu'r gwasanaeth o'r un safon yn y Gymraeg a'r hyn a geir yn Saesneg.

Er gwaethaf pwysigrwydd darparu Cymraeg o'r ansawdd gorau yng Nghymru o fewn modelau iaith mawr, credwn ei bod yn anorfod y bydd yn rhaid mabwysiadu elfen bragmatig wrth geisio sicrhau y nod hwn. Yn ymarferol, ni ellir rhwystro'r defnydd o AI sydd eisoes yn datblygu yn y sector gyhoeddus yng Nghymru, dim ond pwysleisio cyfrifoldebau'r sector, amlygu unrhyw anallu i gwrdd â'r cyfrifoldeb hwnnw, gosod ffordd ymlaen i wneud hynny a'u hatgoffa o'u cyfrifoldebau o ran y safonau iaith.

6 CASGLIADAU

Cyflwynodd y bennod hon y gwerthusiad cynhwysfawr cyntaf o alluoedd Cymraeg modelau iaith mawr OpenAI GPT-3.5 a GPT-4. Trwy gyfres o werthusiadau unigol, aethom ati i fesur perfformiad y modelau hyn ar draws nifer o ddangosyddion allweddol ar gyfer mesur perfformiad LLMs gyda'r Gymraeg. Er i'n canfyddiadau ddangos bod gan y modelau hyn allu syndod o dda yn y Gymraeg, dangosodd ein gwaith bod diffygion amlwg yn parhau. Ar draws y rhan fwyaf o dasgau, perfformiodd GPT-4 yn well na GPT-3.5. Er hynny, cafodd hyd yn oed GPT-4 drafferth gyda rhai agweddau sylfaenol o'r Gymraeg.

Roedd ein canlyniadau gyda chyfieithu peirianyddol yn nodedig am iddynt ddangos y gallai fersiynau wedi eu mireinio o GPT-3.5 berfformio yn debyg neu'n well na GPT-4 am ffracsiwn o'r gost. Awgryma hynny fod yno lwybr addawol ar gyfer datblygu offer iaith Gymraeg cost-effeithiol drwy greu LLMs parth benodol.

Mae ein gwaith yn tanlinellu'r angen i barhau i ddatblygu gwerthusiadau o fodelau iaith yn benodol ar gyfer anghenion Cymru. Bydd y rhain yn hanfodol nid yn unig ar gyfer mesur cynnydd yn y maes, ond hefyd ar gyfer tynnu sylw datblygwyr at yr union agweddau sydd angen eu gwella. Dadleuwn fod angen i'r gwerthusiadau a grëir ar gyfer y Gymraeg hefyd fod yn agored, safonol a throsglwyddadwy fel bod modd eu cymhwyso i nifer o LLMs gwahanol.

Pwysleisiwn hefyd bwysigrwydd rhannu rhagor o ddata o dan drwyddedau agored, yn enwedig o du cyrff cyhoeddus yng Nghymru. Bydd rhannu data o'r fath yn hanfodol er mwyn gwella cynrychiolaeth y Gymraeg o fewn LLMs, a gallai ffurfio ecosystem sy'n fuddiol i'r ddwy ochr lle byddai cyfraniad cyrff cyhoeddus yn gwella modelau a fyddai yn eu tro eu cynorthwyo i ddarparu gwasanaethau Cymraeg yn unol â'r disgwyliadau cyfreithiol sydd arnynt.

Yn olaf, galwn am ddatblygu safonau iaith AI cenedlaethol yng Nghymru ar sail hawliau ieithyddol siaradwyr Cymraeg a'r angen sicrhau cydraddoldeb o ran ansawdd y gwasanaethau AI yn y Gymraeg ac yn Saesneg.

Er yr addewid y gallai AI wella gwasanaethau Cymraeg, mae angen gwaith sylweddol er mwyn sicrhau y bydd y technolegau hyn yn diwallu anghenion unigryw y cyd-destun Cymreig, yn ogystal â'i gofynion cyfreithiol penodol. Bydd gwaith ymchwil, gwerthuso a datblygu polisi parhaus yn hanfodol i wireddu potensial LLMs ar gyfer y Gymraeg, ac i sicrhau bod safonau uchel o ran cywirdeb ieithyddol a phriodoldeb diwylliannol yn cael eu cynnal yn y dyfodol.

DIOLCHIADAU

Diolchwn i Lywodraeth Cymru am ariannu'r gwaith hwn ac i OpenAI am ddarparu'r credydau ar gyfer galwadau API y gwerthusiadau inni am ddim.

CYFEIRIADAU

[1] A. Radford a K. Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. Adalwyd o https://api.semanticscholar.org/CorpusID:49313245

[2] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike a Ryan Lowe. 2022. Training language models to follow instructions with human feedback. Maw. 4, 2022. arXiv:2203.02155.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever a Dario Amodei. 2020. Language Models are Few-Shot Learners. Gor. 22, 2020. arXiv:2005.14165.

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song a Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. Ion. 12, 2021. arXiv:2009.03300.

[5] OpenAI et al. GPT-4 Technical Report. Maw. 4 2024. arXiv:2303.08774.

[6] Delyth Prys, Dewi Bryn Jones, Gruffudd Prys, a Gareth Watkins. 2023. Lecsicon Cymraeg Bangor Welsh Lexicon version 23.10. Adalwyd o https://github.com/techiaith/lecsicon-cymraeg-bangor/releases/tag/23.10.

[7] Gruffudd Prys. 2023. techiaith/word2vec-cy: Model Iaith Fectorau Cymraeg // Welsh Word2Vec Language Model version 0.3. Adalwyd o https://github.com/techiaith/word2vec-cy/releases/tag/v0.3

[8] Delyth Prys, Gruffudd Prys a Dewi Bryn 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.

[9] 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.

[10] Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify a Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. Chw. 17, 2023. arXiv:2302.09210.

[11] H. Xu, Y. J. Kim, A. Sharaf, a H. H. Awadalla. 2024. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. Chw. 6 2024. arXiv:2309.11674.

[12] Senedd y Deyrnas Unedig 1993. Welsh Language Act 1993. Adalwyd o https://www.legislation.gov.uk/cy/ukpga/1993/38/contents

[13] Cynulliad Cenedlaethol Cymru. 2011. Mesur y Gymraeg (Cymru) 2011. Adalwyd o https://www.legislation.gov.uk/mwa/2011/1/contents/enacted/welsh?view=plain

Datblygu offer iaith ar gyfer plant Gwyddeleg eu hiaith sydd ag anghenion ychwanegol

EMILY BARNES

Coleg y Drindod Dulyn

Mae'r bennod hon yn disgrifio'r rhesymeg a'r dull o fynd o'i chwmpas hi i ddatblygu offer iaith ar gyfer plant sy'n siarad Gwyddeleg ac sydd ag anghenion ychwanegol. Byddaf yn canolbwyntio'n arbennig ar ddatblygu system Cyfathrebu Estynedig ac Amgen (Augmentative and Alternative Communication: AAC) mewn Gwyddeleg, gwaith sy'n digwydd yn Labordy Ffoneteg ac Iaith, Coleg y Drindod Dulyn. Mae'r labordy hwn yn lletya project ABAIR, a cheir mwy o wybodaeth am sgôp y datblygiadau technolegol yn y labordy mewn papurau diweddar gan Ní Chasaide a chydweithwyr [1] a Ní Chiaráin a chydweithwyr [2].

Allweddeiriau: Cyfathrebu Estynedig ac Amgen (AAC), Gwyddeleg

1 CYFLWYNIAD

Mae'r system Cyfathrebu Estynedig ac Amgen (Augmentative and Alternative Communication: AAC) yn ap sy'n galluogi defnyddwyr i ddewis cyfres o eiriau neu symbolau sydd wedyn yn cael eu cydgysylltu yn frawddeg ac yn cael eu llefaru gan lais synthetig. Maent yn aml yn cael eu defnyddio gan bobl awtistig nad ydynt yn medru siarad neu heb fedru siarad llawer, yn ogystal â phobl sydd ag anawsterau cyfathrebu. Enw'r system AAC rydym yn ei datblygu yw *Geabaire*, sy'n golygu clebran mewn Gwyddeleg. Ceir adroddiad manylach ar y system yn Barnes et al. [3].

Mae nifer o heriau ynghylch datblygu technoleg mewn iaith leiafrifol. O'r safbwynt ideolegol ceir heriau yn ymwneud ag unieithrwydd a thueddiadau eingl-ganolog, edrych ar anabledd fel diffyg, a'r olwg iwtilitaraidd ar ddefnyddioldeb ieithoedd lleiafrifol [4]. O safbwynt ymarferol, rydym yn dod ar draws heriau o ran gwahaniaethau ieithyddol a sosioieithyddol rhwng Gwyddeleg a Saesneg. Edrychir ar y rhain yn yr adrannau dilynol.

2 HERIAU IDEOLEGOL

Nid yw technolegau cynorthwyol mewn Gwyddeleg eto mor ddatblygedig ag ydynt yn Saesneg, er bod cynnydd sylweddol wedi'i wneud yn ystod y blynyddoedd diwethaf. Wrth gwrs, sylfaen datblygiad technoleg gynorthwyol yw'r canfyddiad fod yno boblogaeth o bobl sydd ei angen. Yn hanesyddol mae plant sydd ag anghenion ychwanegol wedi cael eu hannog i beidio mynychu ysgolion cyfrwng Gwyddeleg, a dyna'r sefyllfa o hyd (e.e. [5]). Yn yr adrannau dilynol, rwy'n trafod y rhesymau ideolegol dros geisio cynghori plant gydag anghenion ychwanegol rhag dilyn addysg drochi neu addysg iaith.

Er gwaetha'r canfyddiadau hyn, mae plant ag anghenion ychwanegol yn mynychu ac yn llwyddo mewn ysgolion cyfrwng Gwyddeleg (e.e. [5]). Yn wir, mae ymchwil yn dangos nad oes unrhyw dystiolaeth dros beidio cynghori plant i fynd i ysgolion dwyieithog neu addysg drochi [6], a bod atal plant rhag dod yn ddwyieithog yn cyfyngu ar eu datblygiad cymdeithasol a diwylliannol [7].

2.1 Tuedd tuag at unieithrwydd ac agwedd eingl-ganolog

Gall y siaradwr uniaith Saesneg gael ei weld fel y status quo mewn gwledydd lle mai Saesneg yw'r brif iaith i fwyafrif y bobl, fel yn Iwerddon. Mae hyn yn adlewyrchu tueddiadau tuag at unieithrwydd, sy'n:

> "assumes that humans have the capacity to acquire one language completely and without difficulty and that acquisition of additional languages is cognitively challenging and often results in incomplete mastery" [8].

Mae hyn yn cyd-fynd gyda bod yn eingl-ganolog, sydd yn:

> "nothing special at all, and at the same time, it is something truly exceptional. It involves a very common cognitive bias called ethnocentrism, in which the norms, values, and repertories of meaning belonging to a specific group, are imposed on other people") [9].

Gyda'i gilydd mae'r tueddiadau hyn yn amlygu eu hunain mewn persbectifau mai unieithrwydd yw'r norm ac - o fewn hynny – mai Saesneg yw'r status quo.

Er mai ideolegol yw'r tueddiadau hynny, mae iddynt oblygiadau ymarferol. Un o'r rhain – sy'n berthnasol i'r drafodaeth am adnoddau i blant gydag anghenion ychwanegol – yw'r diffyg offer asesu i blant mewn ysgolion cyfrwng Gwyddeleg ac yn y Gaeltacht. Cafodd hyn ei adlewyrchu ym marn seicolegydd addysg yn y Gaeltacht mewn cyfweliad nifer o flynyddoedd yn ôl:

> "Má théann tú ar ais agus má dhéanann tú ailinís ar an gcúis a bhfuil an meon sin ann, is é mo bharúil é, ní thuigeann agus ní fheiceann agus ní ghlacann daoine leis go dtagann páistí isteach geata na scoile sna naoíonáin bheaga agus gan acu ach Gaeilge agus níl Béarla acu." [10].

Cyfieithiad:

> "Os ewch chi'n ôl a dadansoddi pam fod y persbectif hwn yn bodoli, fy marn i yw nad yw pobl yn deall nac yn derbyn fod plant yn dod drwy gatiau'r ysgol yn y Babanod Iau [blwyddyn gyntaf yr ysgol] yn gwybod dim byd ond Gwyddeleg, ac nad ydynt yn gwybod Saesneg."

Mae hyn yn tanlinellu'r duedd tuag at Saesneg fel iaith gyntaf, a diffyg cydnabyddiaeth – ac ystyriaeth – o'r Wyddeleg fel iaith gyntaf.

2.2 Addysg iaith myfyrwyr gydag anghenion ychwanegol: golwg diffyg

Mae problem ymyleiddio yn arbennig o eglur yn achos plant ag anghenion ychwanegol. Mae'r polisi a'r cyd-destun deddfwriaethol yn awgrymu nad yw plant gydag anghenion ychwanegol yn cael eu gweld i fod yn byw yn y gofod iaith nac addysg drochi. Yn Iwerddon, mae'r holl blant sydd mewn ysgol arbennig neu sy'n mynychu dosbarth arbennig mewn ysgol prif lif yn awtomatig yn cael eu hesgusodi o wersi Gwyddeleg. Gall myfyrwyr sydd â chyrhaeddiad dysgu isel mewn Saesneg hefyd wneud cais i gael eu hesgusodi [11].

Dadleuir yma fod hyn yn adlewyrchu golwg diffyg ar anabledd. Mae'r olwg diffyg yn proffilio pobl o ran eu hanabledd, gan gymryd y ffocws i ffwrdd oddi wrth eu galluoedd, cryfderau ac anghenion [12]. Mae'r olwg diffyg – yn arbennig yng ngoleuni tueddiadau unieithrwydd ac eingl-ganolog – yn cysyniadu dwyieithrwydd ac amlieithrwydd fel pethau sydd y tu hwnt i bobl sydd ag anghenion ychwanegol neu sydd yn anabl.

2.3 Defnyddioldeb canfyddedig y Wyddeleg: esgusodi rhag y Wyddeleg mewn addysg cyfrwng Saesneg

Mae'n arwyddocaol mai'r Wyddeleg yw'r unig faes addysgol lle cynigir esgusodi plant sydd ag anghenion addysgol ychwanegol mewn ysgolion cyfrwng Saesneg. Yn Iwerddon, mae myfyrwyr yn gweld y Wyddeleg fel peth llai

defnyddiol ond anoddach na phynciau eraill [13]. Rydym yn ffocysu fwyfwy ar addysg fel modd o ffurfio cyfalaf dynol yn yr amseroedd presennol. Gwelir addysg ffurfiol fel modd o ffurfio cyfalaf dynol lle caiff myfyrwyr y sgiliau i ddod yn rhan o'r gweithlu (e.e. [14]). Mae hyn yn cyferbynnu gyda ffocws ar ddatblygiad holistig yr unigolyn, yn academaidd, cymdeithasol, emosiynol a diwylliannol.

2.4 Crynodeb

Er bod cynnydd sylweddol wedi'i wneud yn y blynyddoedd diweddar, nid yw datblygiad technoleg gynorthwyol ar gyfer y Wyddeleg mor ddatblygedig ag ydyw ar gyfer y Saesneg. Yma, cynigir nifer o resymau ideolegol sy'n cyd-gyffwrdd am y diffyg ffocws hanesyddol ar ddatblygiad technoleg gynorthwyol ar gyfer y Wyddeleg. Maent yn cynnwys:

1. y canfyddiad fod dwyieithrwydd neu amlieithrwydd yn heriol yn wybyddol
2. y syniad fod myfyrwyr sydd ag anghenion ychwanegol yn llai abl i ddysgu iaith, a
3. llai o gymhelliad i neilltuo adnoddau i ddysgu iaith leiafrifol oherwydd y canfyddiad o ddiffyg defnyddioldeb dysgu Gwyddeleg.

Mae hyn yn gwrthgyferbynnu gyda chyflwr presennol gwybodaeth; mae myfyrwyr gydag anghenion ychwanegol yn medru mynychu addysg drochi ac yn wir yn gwneud hynny, ac yn llwyddo i ddod yn ddwyieithog. Gyda hynny mewn golwg, mae'r project ABAIR yng Ngholeg y Drindod Dulyn wedi datblygu cyfres o dechnolegau cynorthwyol. Mae'r adran nesaf yn ffocysu ar y datblygiad diweddaraf, Geabaire.

3 DATBLYGU TECHNOLEG AR GYFER PLANT SY'N SIARAD GWYDDELEG

Mae Geabaire, y system AAC gyntaf mewn Gwyddeleg, yn cael ei chynnal gan nifer o werthoedd craidd, a'i chefnogi gan arbenigedd rhyngddisgyblaethol.

3.1 Gwerthoedd craidd

Mae gwerthoedd yn greiddiol i ddatblygu technoleg, er nad ydynt o hyd yn cael eu mynegi mewn cymaint eiriau. Maent i'w gweld yn y technolegau rydym yn dewis eu datblygu, y graddau rydym yn dewis cynnwys defnyddwyr, yr ieithoedd a'r tafodieithoedd rydym yn darparu, a'r pris rydym yn ei godi amdanynt. Gall technolegau alluogi neu analluogi, ac fel y dywed Verbeek [15], "when technologies coshape human actions, they give material answers to the ethical question of how to act. This implies that engineers are doing 'ethics by other means': they materialize morality."

Arweinir datblygiad ein system AAC gan bersbectif niwroamrywiaeth, sy'n gweld anabledd fel rhan naturiol o amrywiaeth y ddynoliaeth, sydd ond yn un agwedd ar hunaniaeth ehangach unigolyn, ac un nad oes angen ei 'gwella' [16]. O bersbectif addysg, golyga hyn gefnogi pobl i gymryd mantais o amrediad llawn cyfleoedd addysgol [16]. Credwn y dylai myfyrwyr sydd ddim yn siarad neu ddim yn siarad llawer gael yr un cyfleoedd addysgol â'u cymheiriaid sydd yn siarad, ac mae hynny'n cynnwys addysg iaith.

Gwerth arall sy'n cael ei adlewyrchu yn ein gwaith datblygu yw ffyddlondeb i natur ieithyddol a sosioieithyddol yr iaith Wyddeleg. Er enghraifft, mae'r cynllun tudalen a ddewiswyd gennym yn adlewyrchu cystrawen y Wyddeleg (Berf-Goddrych-Gwrthrych) o'i gyferbynnu â chystrawen y Saesneg sy'n dylanwadu ar gynllun systemau AAC eraill. Mae'r dewisiadau llais synthetig yn adlewyrchu amrywiaeth tafodieithol y Wyddeleg, gyda'r nod o sicrhau fod y llais yn adlewyrchu hunaniaeth y defnyddiwr cyn belled ag y mae modd.

Un arall o'n gwerthoedd sy'n cael eu gyrru gan y gymuned yw ein bod yn dylunio ar y cyd gyda defnyddwyr y dechnoleg gynorthwyol. Yn achos y project AAC, y catalydd oedd mam dau blentyn sy'n defnyddio AAC ac oedd angen y system AAC ar gyfer eu hysgol cyfrwng Gwyddeleg. Ers hynny mae'r fam wedi dod i fod â rhan ganolog yn yr ymchwil, ynghyd â'i phlant a rhwydwaith o ddefnyddwyr AAC sy'n siaradwyr Gwyddeleg.

3.2 Arbenigedd rhyngddisgyblaethol: ieithyddol, sosioieithyddol, addysgol, parth-benodol

Mae angen arbenigedd rhyngddisgyblaethol i ddatblygu technoleg iaith gynorthwyol. Yn ychwanegol at sgiliau datblygu, mae ar y project hwn angen mewnbwn sylweddol oddi wrth ymchwilwyr mewn parthau ieithyddol, sosioieithyddol ac addysgol, yn ogystal ag arbenigedd parth-benodol mewn AAC. Ar hyn o bryd, rydym yn gwerthuso'r fersiwn gychwynnol o Geabaire gyda defnyddwyr AAC yn ogystal ag athrawon, rhieni a therapyddion iaith a lleferydd.

3.2.1 Arbenigedd ieithyddol a sosioieithyddol

Mae dealltwriaeth trylwyr o gyd-destun ieithyddol a sosioieithyddol Iwerddon yn cynnal datblygiad AAC. Enghreifftir dylanwad arbenigedd o'r fath ar benderfyniadau dylunio yn yr enghreifftiau isod:

- Mae nodweddion y system AAC wedi'u dylunio i adlewyrchu nodweddion pwysig o'r Wyddeleg. Enghraifft o hyn yw'r *tobnascanna* (dolenni cyflym) sy'n poblogi ochr dde'r dudalen gartref, yn ogystal â phob tudalen ddilynol. Er enghraifft, mae rhagenwau arddodol (e.e. *le* (gyda) + *mé* (fi) = *liom* (gyda fi)) yn cael eu defnyddio yn aml iawn mewn Gwyddeleg. Er mwyn hwyluso hyn, mae'n hawdd cael at y rhagenwau arddodol ar bod tudalen o'r system AAC (o dan y botwm *réamhfhocail* (arddodiaid)) er mwyn i'r defnyddiwr gael eu dewis yn hawdd.
- Peth arall sy'n dylanwadu ar gynllun y system yw amcangyfrifon o fynychder defnydd geiriau penodol (e.e. [17]). Mae'r botymau ar y bwrdd cartref er enghraifft yn adlewyrchu'r geiriau a ddefnyddir amlaf mewn Gwyddeleg, er enghraifft. Yn ychwanegol, ffurfiau'r ferf ar y dudalen gartref yw'r ffurfiau mwyaf cyffredin o'r ferf honno. Er enghraifft, mae'r ferf *tháinig* (daeth) yn yr amser gorffennol, tra bo'r ferf *éist* (gwrando) yn y modd gorchmynnol.
- Dewiswyd yr eitemau lecsigol i fod yn berthnasol yn ddiwylliannol ac i adlewyrchu amrywiaeth poblogaeth Iwerddon. Er enghraifft mae gennym chwaraeon cenedlaethol Iwerddon – chwarae hyrli a phêl-droed Gwyddelig – yn ogystal â gwyliau traddodiadol ac offerynnau cerdd. Cynhwysir dyddiau gŵyl a geiriau perthnasol o ddiwylliannau eraill hefyd (e.e. Diwali, Eid-al-Fitr a Shogatsu) ar y dudalen achlysuron arbennig.

Bydd arbenigedd ychwanegol mewn synthesis lleferydd yn allweddol i'r cam nesaf o ddatblygu'r system AAC.

- Mae gan y Wyddeleg dair prif tafodiaith a dim safon llafar, a bwriadwn wneud Geabaire ar gael ym mhob tafodiaith cyn y caiff ei ryddhau'n gyhoeddus. Er bod lleisiau ABAIR ar gael yn y tair prif dafodiaith, mae gennym waith ychwanegol i'w gwblhau i sicrhau fod gwahaniaethau lecsigol a gramadegol ym mhob tafodiaith yn cael eu cynnwys.
- Ein nod yw datblygu fersiwn dwyieithog o Geabaire cyn iddo gael ei ryddhau'n gyhoeddus. Yn ogystal â datblygu fersiwn Saesneg o'r ap presennol, bwriadwn alluogi defnyddwyr i symud yn ôl a blaen rhwng y fersiynau Gwyddeleg a Saesneg. Yn y pen-draw, byddai'n dda galluogi cyfnewid cod rhwng Gwyddeleg a Saesneg o fewn ymadrodd neu frawddeg. Rydym yn dilyn yn eiddgar y gwaith cyffrous sy'n digwydd ym Mhrifysgol Bangor (e.e. [18]) yn hyn o beth.

- Tra bo ABAIR yn darparu detholiad o leisiau oedolion, bydd angen lleisiau plant yn y dyfodol agos i sicrhau fod gan ddefnyddwyr fynediad at leisiau sy'n adlewyrchu nodweddion eu hunaniaeth.

Diddorol sylwi fod Canolfan Bedwyr, Prifysgol Bangor eisoes wedi datblygu Technolegau perthnasol a meysydd arbenigedd ar gyfer y Gymraeg (fel y disgrifir yn [19]).

3.2.2 Arbenigedd addysgol

Gan fod ein ffocws ar blant mewn addysg cyfrwng Gwyddeleg ac yn y Gaeltacht, mae datblygu a chynnal cysylltiadau gydag ysgolion yn hanfodol i ni. Cydran allweddol o allbwn project Geabaire yw pecyn hyfforddiant i athrawon i roi arweiniad ar fodelu Cyfathrebu AAC ac ymgorffori'r defnydd o AAC mewn sefyllfa dosbarth cyfan.

3.2.3 Arbenigedd parth-benodol

Mae angen gwybodaeth parth-benodol ym maes datblygu a defnyddio AAC. Er enghraifft, mae'r cynllun symud yn gydran allweddol o Geabaire. Mae cynllun symud yn hwyluso caffael y system AAC mewn ffordd sy'n debyg i'r dull rydyn ni'n dysgu teipio'n rhugl ac yn awtomatig. Yn achos y berfau er enghraifft, cyflwynir pob ffurf berf yn yr un drefn ar bob tudalen. Yn ychwanegol, mae gosodiad y dudalen yn deillio o Allwedd Fitzgerald, sy'n ddull o godio rhannau ymadrodd yn ôl lliw (e.e. mae berfau wedi'u codio mewn gwyrdd).

3.3 Dyluniad gyda'r ffocws ar y defnyddiwr

Cynorthwyodd defnyddwyr AAC gyda datblygiad Geabaire o'r cychwyn. Un o'r ffyrdd mae defnyddwyr wedi cynorthwyo Geabaire yw drwy ddatblygu straeon defnyddwyr (e.e. [20]). Mae straeon datblygwyr yn gysyniad cymharol syml, maent yn cyfleu gofyniad sydd gan y defnyddiwr sy'n cynrychioli nodwedd neu uned o waith ar gyfer datblygwyr. Golyga hyn ymgynghori gyda defnyddwyr, datblygu dealltwriaeth dda o'r math o nodweddion maent eu hangen i wneud y dechnoleg yn addas iddynt. Rydym wedi ymgynghori gyda defnyddwyr, athrawon, a therapyddion iaith a lleferydd ar y math o ofynion mae defnyddwyr AAC a'u partneriaid cyfathrebu eu hangen ar gyfer y project AAC ac mae hyn yn tywys ein gwaith datblygu a'n blaenoriaethau.

Mae gan stori defnyddiwr y strwythur canlynol: Fel <rôl> rwy eisiau <gweithred> er mwyn <gwerth>. Dyma rai enghreifftiau yn ein cyd-destun ni:

- Fel defnyddiwr AAC, rwy eisiau bwrdd rwy'n gallu ei addasu lle medra i roi geiriau rwy'n eu defnyddio'n aml, er mwyn i mi gael mynediad cyflym at bethau sy'n bwysig i mi.
- Fel partner cyfathrebu, rwy eisiau ffwythiant canfod geiriau fydd yn caniatáu i mi gael hyd i air ar y bwrdd, fel mod i'n gallu modelu yn effeithiol ac yn effeithlon i'm plentyn.
- Fel athrawes, rwy eisiau bysellfwrdd gyda llythrennau Gwyddeleg arno o fewn y system AAC, i'm galluogi i ddangos sut mae gair yn cael ei sillafu i fyfyriwr yn fy nosbarth.

Rydym newydd gychwyn astudiaeth lle bydd defnyddwyr AAC sy'n oedolion, athrawon sy'n bartneriaid cyfathrebu i ddefnyddwyr AAC ifanc a therapyddion iaith a lleferydd yn gwerthuso'r system AAC. Ffocws yr ymchwil ar gyfer y defnyddwyr AAC yw eu profiad defnyddiwr cyffredinol, tra bydd athrawon a therapyddion iaith a lleferydd yn ffocysu ar addasrwydd Geabaire yn yr ystafell ddosbarth neu sefyllfa glinigol. Mae prawf canfyddiad, yn archwilio pa mor ddealladwy, naturiol a gramadegol gywir yw Geabaire hefyd yn cael ei gynllunio ar gyfer rhan gyntaf 2024. Tuag at ail hanner eleni, bwriadwn ymchwilio gyda defnyddwyr AAC iau. Mae angen cynllunio meddylgar i sicrhau bod plant sydd ddim yn siarad neu ddim yn siarad llawer ac nad ydynt eto yn ddefnyddwyr

AAC rhugl yn cymryd rhan mewn modd ystyrlon. Ar hyn o bryd rydym yn archwilio dulliau ymchwil trwy gymryd rhan (gweler e.e. [21, 22]) i'r diben hwn.

Mewn unrhyw broject yn ymwneud â thechnoleg, mae penderfyniadau pwysig i'w gwneud o ran diogelu preifatrwydd defnyddwyr. Yn achos Geabaire, byddai'n ddefnyddiol casglu data ar fynychder defnydd geiriau gan ddefnyddwyr er mwyn deall yn well gynllun y dudalen ar gyfer fersiynau diweddarach o'r dyluniad. Fodd bynnag, mae hyn yn peryglu preifatrwydd defnyddwyr gan na fyddant efallai am i ymchwilwyr gael mynediad at gynnwys eu sgyrsiau. Yr ateb ar hyn o bryd – ar ôl trafod gyda defnyddwyr – yw gofyn i ddefnyddwyr a hoffent optio i mewn neu allan o ddarparu gwybodaeth mynychder yn ddi-enw wrth iddynt lwytho'r ap i lawr. Hefyd, bydd gan ddefnyddwyr sy'n optio i mewn fotwm fydd yn eu galluogi i droi i ffwrdd y recordiad mynychder pan fyddant am wneud hynny.

4 CASGLIADAU

I gloi, credwn y dylai pob myfyriwr gael mynediad at eu hiaith frodorol ac at addysg iaith. Ni ddylai mynediad gael ei gyfyngu gan dueddiadau unieithrwydd nac eingl-ganolog, na chan olwg diffyg o anabledd. Fodd bynnag, yng ngeiriau Engstrom and Tinto [23], nid yw mynediad heb gefnogaeth yn gyfle. Ar gyfer y Saesneg, mae cefnogaeth ar ffurf technolegau cynorthwyol yn aml yn cael eu datblygu gan gwmnïau masnachol. Ar gyfer ieithoedd lleiafrifol fel y Wyddeleg a'r Gymraeg, nid yw endidau masnachol efallai â diddordeb mewn datblygu technolegau o'r fath oherwydd maint bychan y farchnad, o'i chymharu â ieithoedd mawr y byd. Mae o'r pwys mwyaf felly fod cyllid cyhoeddus digonol yn cefnogi'r gwaith datblygu cychwynnol ac yn cynnal y technolegau hyn i'r dyfodol, yn ogystal â meithrin meysydd arbenigedd perthnasol. Mae datblygu technoleg gynorthwyol angen arbenigedd helaeth nid yn unig mewn rhaglennu a synthesis lleferydd, ond hefyd yn y parthau ieithyddol, sosioieithyddol ac addysgol. Mae'n dasg rhyngddisgyblaethol sydd yn cael ei gyrru a'i gweithredu orau gan y cymunedau iaith maent yn eu gwasanaethu, a chan y bobl o fewn y gymuned iaith fydd yn defnyddio'r dechnoleg yn y pen-draw.

CYFEIRIADAU

[1] Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Harald Berthelsen, Christoph Wendler, Andy Murphy, Emily Barnes a Christer Gobl. 2019. Leveraging phonetic and speech research for Irish language revitalisation and maintenance. Yn The XIX International Congress of Phonetic Sciences. Melbourne, Awstralia, 994-998.

[2] Neasa Ní Chiaráin, Oisín Nolan, Neimhin Robinson Gunning a Madeleine Comtois. 2023. Filling the SLaTE: examining the contribution LLMs can make to Irish iCALL content generation. Yn Proceedings of the 9th Workshop on Speech and Language Technology in Education (SLaTE) Dulyn, Iwerddon, 176-181.

[3] Emily Barnes, Julia Cummins, Rian Errity, Oisín Morrin, Harald Berthelsen, Christoph Wendler, Andy Murphy, Helen Husca, Neasa Ní Chiaráin ac Ailbhe Ní Chasaide. 2023. Geabaire, the First Irish AAC System: Voice as a Vehicle for Change. Yn Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023). Dulyn, Iwerddon, 129-133.

[4] Emily Barnes. 2024. Inclusive and Special Education in English-Medium, Irish-Medium, and Gaeltacht Schools: Policy and Ideology of a Fragmented System. Yn Lidia Mañoso-Pacheco, José Luis Estrada Chichón a Roberto Sánchez-Cabrero (eds), Inclusive Education in Bilingual and Plurilingual Programs. IGI Global. Hershey, UDA. 80 – 95.

[5] Sinéad Nic Aindriú, Pádraig Ó Duibhir a Joseph Travers. 2020. The prevalence and types of special educational needs in Irish immersion primary schools in the Republic of Ireland. European Journal of Special Needs Education, 35(5), 603-619.

[6] Elizabeth Kay-Raining Bird, Fred Genesee a Ludo Verhoeven. 2016. Bilingualism in children with developmental disorders: A narrative review. Journal of Communication Disorders, 63, 1-14.

[7] Gabrielle E. Reimann ac Allison B. Ratto. 2023. Sociocultural influences of professional language recommendations in bilingual families of children with autism spectrum disorder: A narrative review. Translational Issues in Psychological Science, 9(4), 354–363.

[8] Fred Genesee. 2022. The monolingual bias: A critical analysis. Journal of Immersion and Content-Based Language Education, 10(2), 153-181.

[9] Carsten Levisen. 2019. Biases we live by: Anglocentrism in linguistics and cognitive sciences. Language Sciences, 76, 101173.

[10] Emily Barnes. 2017. Dyslexia Assessment and Reading Intervention for Pupils in Irish-Medium Education: Insights into Current Practice and Considerations for Improvement. M. Phil Dissertation, School of Linguistics, Speech and Communication Sciences, Coleg y Drindod Dulyn.

[11] Llywodraeth Iwerddon. 2022. Circular 0054/2022: Exemptions for the Study of Irish – Revising Circular 0052/2019. Adalwyd o https://www.gov.ie/en/circular/28b2b-exemptions-from-the-study-of-irish-primary/

[12] Janette Dinishak. 2022. The deficit view and its critics. Disability Studies Quarterly, 36(4).

[13] Emer Smyth, Allison Dunne, Merike Darmody a Selina McCoy. 2007. Gearing up for the Exam: the Experiences of Junior Certificate Students. ESRI/DES. Dulyn, Iwerddon.

[14] Ian Hardy a Stuart Woodcock. 2015. Inclusive education policies: Discourses of difference, diversity and deficit. International Journal of Inclusive Education, 19(2), 141-164.

[15] Peter-Paul Verbeek. 2006. Materializing morality: Design ethics and technological mediation. Science, Technology, & Human Values, 31(3), 361-380.

[16] Sara M. Acevedo ac Emily A. Nusbaum. 2020. Autism, neurodiversity, and inclusive education. Oxford Research Encyclopedia of Education. Adalwyd o https://doi.org/10.1093/acrefore/9780190264093.013.1260

[17] Breacadh. 2007. Liostaí Bhreacadh: Focail Choitianta sa Ghaeilge. Breacadh.

[18] Stephen Russell, Dewi Bryn Jones a Delyth Prys. 2022. BU-TTS: An Open-Source, Bilingual Welsh-English, Text-to-Speech Corpus. Yn Proceedings of the 4th Celtic Language Technology Workshop within LREC2022. Marseille, Ffrainc, 104-109.

[19] Delyth Prys a Gareth Watkins. 2023. Language Report Welsh. Yn Rehm, G. & Way, A. (eds), European Language Equality: A Strategic Agenda for Digital Language Equality. Springer. 223-226.

[20] Bill Wake. 2003. INVEST in Good Stories, and SMART Tasks. Adalwyd o http://xp123.com/ articles/invest-in-good-stories-and-smart-tasks/

[21] Naomi Winstone, Corinne Huntington, Lisa Goldsack, Elli Kyrou a Lynne Millward. 2014. Eliciting rich dialogue through the use of activity-oriented interviews: Exploring self-identity in autistic young people. Childhood, 21(2), 190-206.

[22] Eija Sevón, Marleena Mustola, Anna Siippainen, a Janniina Vlasov. 2023. Participatory Research Methods with Young Children: A Systematic Literature Review. Educational Review, 1–19.

[23] Cathy Engstrom a Vincent Tinto. 2008. Access without support is not opportunity. Change: The magazine of higher learning, 40(1), 46-50.

Datblygu lleisiau synthetig dwyieithog newydd ar gyfer plant a phobl ifanc Cymru

Cydweithrediad rhwng Gwasanaeth Iechyd Gwladol Cymru, cwmni Cereproc, Caeredin, a Phrifysgol Bangor

MEINIR WILLIAMS

Prifysgol Bangor

DEWI BRYN JONES

Prifysgol Bangor

SARAH COOPER

Prifysgol Bangor

STEFANO GHAZZALI

Prifysgol Bangor

DELYTH PRYS

Prifysgol Bangor

Datblygwyd wyth pâr o leisiau synthetig Cymraeg a Saesneg i'w defnyddio gan blant a phobl ifanc yng Nghymru sydd ag anawsterau lleferydd ac yn defnyddio dulliau Cyfathrebu Estynedig ac Amgen (AAC). Mae'r lleisiau hyn yn cynrychioli acenion de a gogledd Cymru yn Gymraeg a Saesneg. Mae pedwar o'r lleisiau yn rhai benywaidd, a phedwar yn lleisiau gwrywaidd. Defnyddiwyd cwmni masnachol i ganfod a recordio'r talentau llais, a manteisiwyd ar lyfrgell adnoddau technoleg lleferydd Prifysgol Bangor i fireinio'r lleisiau, gan ddatblygu cydrannau newydd lle roedd angen. Cyfrannodd cwmni Cereproc eu harbenigedd yn adeiladu'r lleisiau ac yn arwain y project cyfan. Casglwyd adborth i'r lleisiau gogleddol gan therapyddion iaith a lleferydd, a chyflwynir y canlyniadau cyntaf yma hefyd.

Allweddeiriau: Cymraeg, Testun i Leferydd, Lleisiau Synthetig, Technoleg Estynedig ac Amgen (AAC)

1 CYFLWYNIAD

Byth ers dyddiau cynnar cyfrifiaduron personol, daeth y syniad o greu llais artiffisial o fewn y cyfrifiadur yn un atyniadol. Roedd nifer o resymau dros geisio gwneud hyn, gan gynnwys galluogi defnyddwyr oedd â nam ar eu golwg i ddarllen testun, a galluogi pobl nad oedd yn medru siarad i gyfathrebu ar lafar. Datblygwyd technoleg testun i leferydd yn gyntaf o'r 1960au ymlaen ar gyfer y Saesneg, ac mor ddiweddar â 1990, lleisiau gwrywaidd oedd y rhain bron yn ddieithriad. Yn y flwyddyn honno datblygodd Ann Syrdal y llais benywaidd cyntaf, datblygiad nodedig i gydraddoldeb merched [1]. Yn y cyfnod hwn roedd y syniad o gael technoleg o'r fath yn gweithio ar gyfer y Gymraeg i weld yn bell i ffwrdd, ond yn 2003 cafodd Prifysgol Bangor yng Nghymru a Phrifysgolion Trinity, DCU a UCD yn Iwerddon grant o Gronfa Interreg yr Undeb Ewropeaidd i ddatblygu adnoddau iaith a lleferydd Cymraeg a Gwyddeleg ar y cyd (Welsh and Irish Speech Processing Resources: WISPR). Yr oedd hwn yn broject allweddol roddodd sylfaen ar gyfer ymchwil pellach yn y ddwy iaith. Roedd y lleisiau cyntaf ddatblygwyd yn seiliedig ar dechnoleg deuffonau, ac yn swnio'n robotaidd iawn, ac eto yn ddealladwy ac yn gaffaeliad yn enwedig i bobl oedd â nam ar eu golwg [2].

Ochr yn ochr â'r ymchwil testun i leferydd dechreuwyd datblygu ymchwil adnabod lleferydd ar gyfer y Gymraeg [3], sef lle mae pobl yn siarad gyda'r cyfrifiadur a'r cyfrifiadur naill ai'n ymateb i'r hyn a ddwedwyd e.e. drwy gynnau golau neu chwilio am wybodaeth, neu drosi'r geiriau llafar yn destun, e.e. ar gyfer trawsgrifio neu isdeitlau. Yn y cyfnod hwn roedd gallu technoleg i greu gwell lleisiau synthetig a gwell adnabod lleferydd yn cynyddu'n sydyn, yn enwedig ar ôl dechrau defnyddio dulliau AI a modelau iaith i'w creu. Datblygodd Uned Technolegau Iaith Prifysgol Bangor athroniaeth o gyhoeddi cod cyfrifiadurol ac adnoddau yn deillio o'u hymchwil dan drwyddedau agored caniataol hyd yr oedd modd, ar lwyfannau megis GitHub a'r ELG, gyda'r Porth Technolegau Iaith Cenedlaethol yn gweithredu fel man i gael gwybodaeth amdanynt i gyd [4]. Pwrpas hyn oedd galluogi cwmnïau masnachol bach a mawr, ac ymchwilwyr eraill i fanteisio ar yr adnoddau a'u datblygu ymhellach, gan fod yn ymwybodol na fyddai ieithoedd bach fel y Gymraeg yn cael eu cynnwys mewn rhaglenni technoleg lleferydd oni bai ei bod yn hawdd ac yn rhad cael gafael ynddynt. Comisiynodd yr RNIB gwmni Ivona yng Ngwlad Pwyl, gyda nawdd Llywodraeth Cymru, i greu lleisiau synthetig o ansawdd da i alluogi darllen testun Cymraeg pobl â nam ar eu golwg, a chafwyd llais benywaidd (Gwyneth) a llais gwrywaidd (Geraint) o safon uwch na'r un a gafwyd o'r blaen ar gyfer y Gymraeg [5].

Wrth i'r dechnoleg wella, daeth hi'n bosib adeiladu llais synthetig oedd yn debyg i lais go iawn unigolyn. Cychwynnwyd cynnig adeiladu lleisiau synthetig tebyg i'w llais eu hun i unigolion oedd ar fin colli eu lleferydd oherwydd cyflyrau meddygol, ond dim ond yn Saesneg oedd y gwasanaeth hwn ar gael ym Mhrydain. Un corff felly oedd Cymdeithas Clefyd Niwronau Echddygol (MND) [6] oedd â gwasanaeth bancio lleisiau ond heb yr arbenigedd i ddarparu ar gyfer y Gymraeg. Gyda nawdd Llywodraeth Cymru a chefnogaeth y Gwasanaeth Iechyd Gwladol a nifer o therapyddion iaith a lleferydd, lluniwyd project Lleisiwr i ddatblygu gwasanaeth tebyg ar gyfer y Gymraeg [7]. Oherwydd yr angen i ddiogelu preifatrwydd cleifion a defnyddwyr gwasanaeth, a gochel rhag y perygl o gamddefnyddio lleisiau unigolion, ni ryddhawyd y lleisiau hyn yn gyhoeddus. Yn y cyfamser, daeth hi'n amlwg fod angen lleisiau Cymraeg synthetig ar gyfer plant a phobl ifanc oedd yn defnyddio dyfeisiau cyfathrebu digidol, lle nad oedd hi'n ddelfrydol clywed plant a phobl ifanc yn defnyddio lleisiau oedolion i geisio cyfathrebu ar lafar. O'r profiad yn datblygu Lleisiwr hefyd sylweddolwyd fod unigolion dwyieithog yng Nghymru yn newid iaith yn ôl ac ymlaen drwy'r dydd rhwng y Gymraeg a'r Saesneg, ac y byddai'n rhyfedd petaent yn defnyddio lleisiau gwahanol ar gyfer y ddwy iaith. Canlyniad hyn oedd dechrau ymchwilio i fodelau dwyieithog fyddai'n galluogi defnyddio yr un llais i newid rhwng y ddwy iaith yn ôl y galw, yn cynnwys lleisiau gwirioneddol ddwyieithog a pharau o leisiau.

Yn ogystal, mae systemau Cyfathrebu Estynedig ac Amgen (AAC) wedi datblygu'n sylweddol yn y blynyddoedd diwethaf, ac felly mae'r angen am leisiau soffistigedig i'w defnyddio gyda'r rhain wedi esblygu. Mae AAC yn cynnwys unrhyw ddulliau cyfathrebu sy'n cynorthwyo neu'n disodli iaith lafar. Gall hyn gynnwys ystod eang o ddulliau cyfathrebu, o ystumiau a ieithoedd arwyddo i ddyfeisiadau electronig sy'n cynhyrchu lleferydd, ac mae'r dulliau gwahanol yn cael eu defnyddio yn dibynnu ar anghenion unigolion [8]. Cyn y prosiect hwn, roedd modd i unigolion ddefnyddio llais Gwyneth gyda system Gymraeg elfennol ar feddalwedd Smartbox Grid 3.

2 MANYLION Y PROJECT

Ym mis Medi 2021 cyhoeddodd Gwasanaeth Iechyd Cenedlaethol Cymru hysbysiad yn gofyn am geisiadau i greu lleisiau plant Cymraeg a Saesneg gydag acen Gymreig ar gyfer cymhorthion cyfathrebu technolegol. Mewn ymateb i'r hysbysiad, penderfynodd cwmni CereProc o Gaeredin a Phrifysgol Bangor gydweithio er mwyn cyflwyno tendr ar y cyd. Fel uned ymchwil brifysgol, mae pwyslais yr Uned Technolegau Iaith yn y brifysgol ar ddatblygu'r ymchwil angenrheidiol, heb gystadlu yn erbyn y sector breifat oni bai fod methiant yn y farchnad, a chan gydweithio gyda diwydiant er lles cymdeithas a'r economi. Mae CereProc yn gwmni testun i leferydd arbenigol yn yr Alban sy'n

arwain y byd ar dechnoleg o'r fath. Roedd hi'n bartneriaeth addas iawn felly gyda Cereproc yn cyfrannu eu harbenigedd technolegol a masnachol, a'r brifysgol yn cyfrannu eu harbenigedd o dechnoleg lleferydd Gymraeg a materion ieithyddol a ffonolegol. Yr oedd Prifysgol Bangor eisoes wedi rhyddhau corpws testun i leferydd o leisiau oedolion yn Gymraeg [9], ac wedi hyfforddi llais niwral oedd ar gael dan drwydded agored ar y we [10] ac yn awyddus iawn i gydweithio gyda CereProc i gynhyrchu lleisiau o safon masnachol uchel. Dechreuwyd y prosiect ym mis Ionawr 2023 a gorffennwyd y lleisiau ar ddiwedd mis Medi yr un flwyddyn.

Datblygwyd 8 pâr o leisiau testun i leferydd Cymraeg a Saesneg ar gyfer plant a phobl ifanc mewn dwy dafodiaith – de a gogledd Cymru. Teimlwyd bod hyn yn angenrheidiol o ystyried nad oes un dafodiaith safonol yn y Gymraeg ac bod rhoi llais cyn debyced â phosib i'w cyfoedion i'r defnyddwyr yn un o nodau'r prosiect. Roedd pob pâr wedi eu lleisio gan yr un talent llais er mwyn sicrhau bod y defnyddwyr yn gallu cadw'r un llais wrth siarad Cymraeg a Saesneg – ystyriaeth bwysig mewn gwlad ddwyieithog. Nid yw'r lleisiau eu hunain yn ddwyieithog, ond mae peth cyfnewid côd yn bosib. Golyga hyn bod modd i unigolion ddefnyddio'r ddwy iaith mewn modd llawer mwy integredig yn hytrach na'u cadw nhw ar wahân.

Roedd angen datblygu lleisiau ar gyfer ystod o oedrannau gwahanol er mwyn adlewyrchu lleisiau plant a phobl ifanc, sy'n datblygu'n sylweddol. Penderfynwyd datblygu 4 llais ar gyfer plant 8-12 oed, a fyddai hefyd yn addas ar gyfer plant iau, a 4 llais ar gyfer pobl ifanc yn eu harddegau (13-16+), a fyddai'n debycach i leisiau oedolion. Gwelir manylion y lleisiau yn y tabl isod.

Tabl 1: Manylion y lleisiau

Enw	Rhywedd	Acen	Oed
Ffion	Benyw	Gogleddol	13-16+
Seren	Benyw	Gogleddol	8-12
Tomos	Gwryw	Gogleddol	13-16+
Owain	Gwryw	Gogleddol	8-12
Rhian	Benyw	Deheuol	13-16+
Catrin	Benyw	Deheuol	8-12
Rhodri	Gwryw	Deheuol	13-16+
Gethin	Gwryw	Deheuol	8-12

Er mwyn creu'r lleisiau, roedd angen dod o hyd i dalentau llais a fyddai'n adlewyrchu'r grwpiau oedran ac acenion daearyddol yn y Gymraeg a'r Saesneg. Bu cwmni cynhyrchu teledu Darlun yn gwrando ar dros 300 o blant o oedrannau gwahanol cyn anfon samplau o 16 llais i'r tîm ym Mangor a Cereproc i wneud y penderfyniad terfynol. Er mwyn gwneud hyn recordiwyd y talent yn darllen sgript 200 o frawddegau a oedd wedi eu cynllunio i gynnwys holl ffonemau'r Gymraeg a'r Saesneg. Roedd hyn yn seiliedig ar Eiriadur Ynganu Bangor [11], a addaswyd ar gyfer y prosiect hwn, a sgript Cereproc a oedd yn bodoli eisoes ar gyfer lleisiau Saesneg. Paratowyd sgriptiau Cymraeg a Saesneg o tua 1,200 brawddeg yr un ar gyfer recordio'r lleisiau, yn cynnwys adrannau a oedd yn adlewyrchu'r acenion gogleddol a deheuol. Roedd rhaid sicrhau bod digon o esiamplau o bob ffonem yn y sgript er mwyn adeiladu'r lleisiau, a defnyddiwyd nofelau i blant (o fewn rheolau hawlfraint) i sicrhau bod y deunydd ar lefel resymol i'r talent ei ddarllen. Unwaith roedd y recordio wedi ei wneud, adeiladwyd y fersiwn gyntaf o'r lleisiau gan Cereproc.

Yn dilyn hyn, dechreuwyd ar y broses o normaleiddio'r lleisiau. Gan fod Cereproc eisoes wedi adeiladu nifer o leisiau Saesneg roedd y broses yn gymharol syml ar gyfer y lleisiau Saesneg, ond roedd rhaid datblygu normaleiddiwr ar gyfer rhifau, unedau, amseroedd ayb yn y Gymraeg. Defnyddiwyd normaleiddiwr a oedd wedi ei greu eisoes gan yr Uned Technolegau Iaith [12] fel sail. Roedd nifer o broblemau i'w hystyried yn ymwneud â'r rhifau, yn enwedig defnyddio'r systemau degol ac ugeiniol mewn cyd-destunau gwahanol. Penderfynwyd bod defnyddio'r system ddegol (11 = un deg un, 12 = un deg dau, ayb.) yn fwy addas ar gyfer defnydd cyffredinol gan fod y lleisiau wedi eu creu ar gyfer plant, tra bod y system ugeiniol yn cael ei chadw at ddyddiadau, e.e. y pymthegfed, ac amseroedd e.e. ugain munud wedi tri, yn unig. Roedd ynganiad byrfoddau hefyd yn destun ystyriaeth gan fod rhai yn defnyddio ynganiad Cymraeg, e.e. S4C, eraill yn cael eu hynganu fel gair, e.e. CBAC, rhai'n cael eu hynganu yn defnyddio enwau llythrennau Saesneg, e.e. BBC.

Ystyriaeth bwysig arall oedd gallu'r lleisiau i ymdopi â chyfnewid côd gan fod hynny'n ffenomen cyffredin iawn mewn Cymraeg cyfoes [13]. Roedd hefyd yn hanfodol bod y lleisiau Saesneg yn gallu ymdopi â rhai geiriau Cymraeg, yn enwedig enwau llefydd. Er mwyn gwneud hyn addaswyd geiriadur ynganu Saesneg Cereproc i gynnwys ffonemau'r Gymraeg ac fe gyflwynwyd y rheolau ynganu Cymraeg i'r lleisiau Saesneg a'r gwrthwyneb. Fodd bynnag, mewn enghreifftiau lle'r oedd rheolau orgraff y ddwy iaith yn cyferbynnu e.e. *Allan* yn Saesneg /ælən/ ac *allan* yn Gymraeg /ałan/, cadwyd at brif iaith pob llais er mwyn sicrhau cysondeb.

Mae'r lleisiau ar gael i'w lawrlwytho apiau CereProc Voices ar systemau Windows a MacOS, ac ar iOS ac Android. Byddant yn addas i'w defnyddio gyda nifer o feddalweddau AAC, yn cynnwys Smartbox, Liberator, Tobii Dynavox a Jabbla. Gobeithir cydweithio'n ehangach yn y dyfodol i sicrhau bod y lleisiau ar gael i unrhyw un a fyddai'n cael budd o'u defnyddio.

3 CANLYNIADAU

Cwblhawyd yr 8 pâr o leisiau ac mae'r broses o'u lansio a gwerthuso'n parhau er mwyn sicrhau bod y lleisiau'n addas ar gyfer eu defnyddio yn y cyd-destunau a fwriedid.

Cyflwynwyd fersiwn beta o'r lleisiau gogleddol Cymraeg a Saesneg i grŵp o therapyddion iaith a lleferydd, ymchwilwyr a myfyrwyr sydd â diddordeb mewn lleferydd ac iaith yn ystod Cyfnewidfa Iaith a Lleferydd Gogledd Cymru (NWSLE) ym mis Hydref 2023.

Cwblhaodd 33 o gyfranogwyr holiadur dilynol ar ôl y sesiwn. Roedd 17 yn Therapyddion Iaith a Lleferydd, 1 yn gynorthwyydd iechyd ac 15 naill ai yn astudio therapi iaith a lleferydd, yn fyfyrwyr ymchwil neu cynorthwyo Therapyddion Iaith a Lleferydd.

O'r rhain, roedd 26 yn fenywaidd, 2 yn wrywaidd, 3 yn Anneuaidd a 2 yn dewis peidio â datgelu eu rhywedd. Roedd 7 o'r cyfranogwyr rhwng 18 ac 24 mlwydd oed, 7 rhwng 25 a 34, 14 rhwng 35 a 44, a 4 rhwng 45 a 54. Roedd 13 yn siarad Cymraeg yn rhugl, 7 yn siarad cryn dipyn o Gymraeg, 5 yn siarad ychydig bach o Gymraeg a 7 yn gallu deud ychydig eiriau. Nododd 16 eu bod yn gweithio yn ngogledd orllewin Cymru, 8 yng nghanol gogledd Cymru, 6 yn y gogledd ddwyrain, a 2 yn gweithio tu hwnt i ogledd Cymru.

Yn ystod y sesiwn, cafwyd trafodaeth am AAC o fewn y bwrdd iechyd lleol (Betsi Cadwaladr) a'r gwasanaeth Technoleg Gynorthwyol Electronig (EAT) ac astudiaeth achos o dechnoleg AAC yn cael ei gweithredu trwy gyfrwng y Gymraeg cyn arddangos y lleisiau.

Cyflwynwyd samplau o'r lleisiau yn siarad Saesneg (yn cynnwys enwau lleoedd Cymraeg), Cymraeg ffurfiol yn cynnwys rhifau, a Chymraeg anffurfiol yn cynnwys cyfnewid côd ac orgraff ansafonol.

Tabl 2: Samplau lleisiau a gyflwynwyd i gynulleidfa'r Gyfnwidfa Iaith a Lleferydd Gogledd Cymru

Llais	Sampl Saesneg	Sampl Cymraeg	Sampl Cymraeg anffurfiol / cyfnewid cod
Ffion	Hello, my name is Ffion and I come from Pwllheli. I am a digital voice for AAC	Ni ydi'r lleisiau digidol sydd wedi cael eu datblygu gan Cereproc a Phrifysgol Bangor	Tisio panad?
Seren	Hello, my name is Seren and I come from Porthmadog.	Mae 4 llais o'r gogledd a 4 llais o'r de	Dani di bod yn Chester Zoo
Tomos	Hello, my name is Tomos and I come from Harlech.	Dwi'n un o'r lleisiau ar gyfer pobl ifanc 13-16 oed	O'dd o'n briliant
Owain	Hello, my name is Owain and I come from Wrexham.	A dwi'n un o'r lleisiau ar gyfer plant 8-12 oed	Ond dwi isio mynd adra

Yn dilyn hyn, gofynnwyd i'r cyfranogwyr lenwi holiadur yn ymateb i'r lleisiau a'r sesiwn yn fwy cyffredinol. Yma, rydym yn cyflwyno'r ymatebion y gynulleidfa i'r lleisiau. Gofynnwyd iddynt ymateb i bum datganiad a dweud i ba raddau roeddent yn cytuno neu'n anghytuno â nhw ar raddfa o 1 (anghytuno'n gryf) a 5 (cytuno'n gryf):

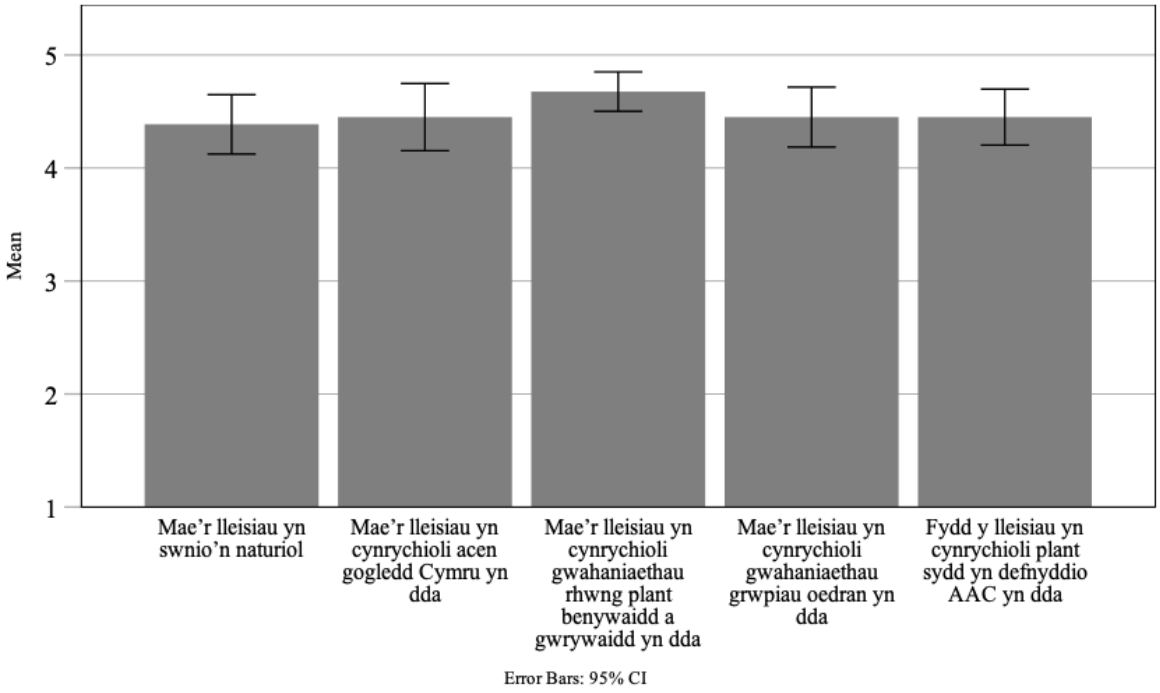
- Mae'r lleisiau yn swnio'n naturiol
- Mae'r lleisiau yn cynrychioli acen gogledd Cymru yn dda
- Mae'r lleisiau yn cynrychioli gwahaniaethau rhwng plant benywaidd a gwrywaidd yn dda
- Mae'r lleisiau yn cynrychioli gwahaniaethau grwpiau oedran yn dda
- Fe fydd y lleisiau yn cynrychioli plant sydd yn defnyddio AAC yn dda

Mae'r canlyniadau i'w gweld yn Ffigur 1. Gofynnwyd i'r cyfranogwyr hefyd lenwi cwestiwn sy'n eu holi am ddatblygiadau yn y maes yn y dyfodol a'r hyn yr oeddent yn meddwl oedd ei angen.

Yn gyffredinol, roedd yr ymateb yn ffafriol iawn, gyda'r mwyafrif o'r ymatebion naill ai'n cytuno (4) neu'n cytuno'n gryf (5) gyda'r datganiadau. Cafwyd rhai sylwadau yn nodi nad oedd y lleisiau'n adlewyrchu acenion y gogledd ddwyrain cystal, sydd yn debygol o fod yn gysylltiedig â'r ffaith bod y talent llais ar gyfer y lleisiau gogleddol yn dod o'r gogledd orllewin. Yn ogystal, nodwyd bod Ffion, y llais benywaidd hŷn, yn ymddangos fel oedolyn, tra bod un arall o'r ymatebion yn nodi bod Tomos, y llais gwrywaidd hŷn, yn uchel o ran traw, ac felly'n ymddangos yn iau.

Nododd sawl cyfranogwr bod angen datblygiadau pellach o ran meddalwedd AAC cyfrwng Cymraeg i'w defnyddio gyda'r lleisiau, yn enwedig o ran gwahaniaethau tafodieithol o ran geirfa, e.e. *llefrith/llaeth*.

Yn ogystal, cafwyd ymateb gan fam merch 6 oed sydd wedi dechrau defnyddio un o'r lleisiau Cymraeg. Roedd yr ymateb i'r llais ei hun yn gadarnhaol, gan bwysleisio mor bwysig oedd cael llais addas ar gyfer oed ei merch "I'm so thrilled her voice is age appropriate now", ond roedd hefyd yn pwysleisio'r ffaith bod y lleisiau'n bodoli mewn system AAC ehangach a bod angen datblygu gweddill y ddarpariaeth cyfrwng Cymraeg i alluogi unigolion i gyfathrebu'n rhwydd yn yr iaith a defnyddio'r lleisiau gymaint â phosib.



Ffigur 1: Ymatebion i holiadur y cyfranogwyr

4 CASGLIADAU

Nod y prosiect hwn oedd creu wyth pâr o leisiau testun i leferydd Cymraeg a Saesneg a fyddai'n addas ar gyfer eu defnyddio gyda systemau AAC o fewn naw mis, ac fe gwblhawyd y nod hwn. Mae llwyddo i greu parau o leisiau mewn iaith leiafrifol (Cymraeg) a iaith fwyafrifol (Saesneg) yn defnyddio'r un talent llais a sgriptiau cyfatebol, yn cynnig y posibilrwydd o ddefnyddio AAC yn ddwyieithog heb orfod newid llais yn gosod patrwm i ieithoedd lleiafrifol eraill eu dilyn.

Mae'r ymateb cychwynnol i fersiwn beta wedi bod yn ffafriol iawn ar y cyfan, ac fe fydd rhyddhau'r lleisiau'n golygu adborth ehangach gan ddefnyddwyr o bob rhan o Gymru. Yn sgil hyn bydd modd ystyried pa agweddau sydd wedi diwallu anghenion defnyddwyr a pha agweddau sydd angen eu datblygu ymhellach.

Efallai mai'r brif ystyriaeth ar gyfer datblygu'r lleisiau'n ehangach ydi datblygu'r feddalwedd y mae'r lleisiau'n gweithredu ynddi. Byddai datblygu tudalennau AAC Cymraeg yn galluogi defnyddwyr i addasu'r lleisiau i adlewyrchu eu dymuniadau nhw o ran cyfathrebu yn cynnwys eu hacenion, geirfa, a ffyrdd o gyfnewid côd nhw eu hunain. Heb y gwaith pellach, mae'r lleisiau'n gam pwysig yn y cyfeiriad cywir, ond nid yw eu potensial llawn yn cael ei wireddu. Rhaid datblygu'r feddalwedd cyfrwng Cymraeg ar gyfer defnyddio'r lleisiau er mwyn sicrhau bod cyfleoedd cyfartal i unigolion sy'n dymuno byw eu bywydau'n ddwyieithog trwy gyfrwng y Gymraeg a'r Saesneg. Mae prosiect arloesol yn datblygu system AAC ar gyfer Gwyddeleg[1] yn dangos cyfeiriad posib i system Gymraeg ei ddilyn, yn enwedig o ystyried yr heriau cyffredin rhwng yr ieithoedd Celtaidd, yn cynnwys treiglo, trefn brawddegau, cymhlethdodau ffurfiau ie/na a rhedeg arddodiaid.

[1] Gweler y bennod flaenorol yn y llyfr hwn.

Mae'r prosiect hwn wedi dangos bod datblygu adnoddau safonol ar gyfer AAC trwy gyfrwng dwy iaith yn bosib, ac yn ateb posib i ieithoedd bychain sydd yng nghysgod ieithoedd mwy mewn cymunedau dwyieithog. Gobeithir y bydd y patrwm hwn yn gynsail i ddatblygu adnoddau i sicrhau bod ystod ehangach o ieithoedd ar gael i bawb sy'n dymuno eu defnyddio.

DIOLCHIADAU

Diolchwn o galon i Cereproc, Darlun, y talentau llais, Gwasanaeth EAT GIG Cymru, Jeff Morris, NWSLE, Rebecca Day a phawb arall sydd wedi cefnogi'r prosiect.

CYFEIRIADAU

[1] Cade Metz. 2020. Ann Syrdal, Who Helped Give Computers a Female Voice, Dies at 74. New York Times (20.08.2020). Adalwyd o https://www.nytimes.com/2020/08/20/technology/ann-syrdal-who-helped-give-computers-a-female-voice-dies-at-74.html

[2] Delyth Prys, Briony Williams et al. 2004. WISPR: Speech Processing Resources for Welsh and Irish. Pre-Conference Workshop on First Steps for Language Documentation of Minority Languages, LREC Conference, Lisbon, Portiwgal.

[3] Sarah Cooper, Dewi Bryn Jones a Delyth Prys. 2014. Developing further speech recognition resources for Welsh. Yn Judge, J., Lynn, T., Ward, M. and Ó Raghallaigh, B. (eds), Proceedings of the First Celtic Language Technology Workshop at the 25th International Conference on Computational Linguistics (COLING 2014). Dulyn, Iwerddon, 55-59.

[4] Delyth Prys a Dewi Bryn Jones. 2018. National Language Technologies Portals for LRLs: a Case Study. Lecture Notes in Artificial Intelligence. Springer.

[5] RNIB Cymru. Lleisiau synthetig Cymru. Adalwyd o https://www.rnib.org.uk/cy/nations/cymruwales/lleisiau-synthetig-cymru/

[6] Motor Neuron Disease Association. 2022. Voice Banking for Motor Neurone Disease. Adalwyd o https://www.mndassociation.org/sites/default/files/2023-05/P10%20Voice%20banking%202022%20v2.pdf

[7] Bangor University's Language Technologies Unit. 2018. Lleisiwr. Adalwyd o https://lleisiwr.techiaith.cymru/

[8] Kristi L. Morin, Jennifer B. Ganz, Emily V. Gregori, Margaret J. Foster, Stephanie L. Gerow, Derya Genç-Tosun ac Ee Rea Hong. 2018. A systematic quality review of high-tech AAC interventions as an evidence-based practice. Augmentative and Alternative Communication, 34(2), 104–117.

[9] Uned Technolegau Iaith Prifysgol Bangor. 2021. Corpws Talentau Llais. Adalwyd o https://git.techiaith.bangor.ac.uk/data-porth-technolegau-iaith/corpws-talentau-llais

[10] Stephen Russell, Dewi Bryn Jones a Delyth Prys. 2022. BU-TTS: An Open-Source, Bilingual Welsh-English, Text-to-Speech Corpus. In Proceedings of the 4th Celtic Language Technology Workshop within LREC2022. Marseille, Ffrainc, 104-109.

[11] Bangor University's Language Technologies Unit. 2021. Geiriadur Ynganu Bangor | Bangor Pronouncing Dictionary. Retrieved from https://github.com/techiaith/geiriadur-ynganu-bangor/tree/21.03

[12] [9] Uned Technolegau Iaith Prifysgol Bangor. 2023. techiaith-tts. Adalwyd o https://github.com/techiaith/techiaith-tts/tree/main

[13] Margaret Deuchar a Peredur Davies. 2009. Code switching and the future of the Welsh language. International Journal of the Sociology of Language, 195 (Ion. 2009), 15-38.

Adendwm

MÉLANIE JOUITTEAU - L'UNIVERSITÉ BORDEAUX MONTAIGNE A L'UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

Mae deallusrwydd artiffisial ar y ffordd, a rhaid i'n hieithoedd ni fod yn barod

Pwynt syml sydd gen i. Rhaid i'n hieithoedd ni fedru delio gyda deallusrwydd artiffisial os ydyn nhw i oroesi dyfodiad y rhaglenni cyfrifiadurol newydd sy'n seiliedig ar ddata personol.

Mae deallusrwydd artiffisial (AI) yn system sy'n cael ei nodweddu gan fodelau a hyfforddwyd ar setiau data helaeth, gan gynnwys data ieithyddol, ac mae wedi dod yn rhan annatod o'n bywydau bob dydd. Gwelir hyn yn amlwg o'r ffordd rydym yn defnyddio technoleg teipio darogan ac adnabod lleferydd ar ein ffonau clyfar. Dim ond dechrau'r daith yw hyn, ac mae yma debygrwydd rhwng y ffordd mae AI yn dod yn rhan integredig o'n cymdeithas â'r hyn ddigwyddodd gyda ffonau symudol. Mewn dadansoddiad manwl yn cynnwys deuddeg pwynt, mi wna i amlinellu'r ffordd y gwnaeth y chwyldro technolegol blaenorol, sef ffonau symudol, ddod yn rhan integredig o'n ffordd o fyw. Fe fyddaf yn tynnu cymariaethau gyda'r chwyldro AI presennol ym mhob un o'r camau hyn.

Mae camau presennol datblygiad AI yn gosod y sylfaen ar gyfer gweithredu rhaglenni fydd yn ddibynnol ar ddata personol. Cynigiaf esboniad manwl o pam, o'm safbwynt i, mae rhaglenni AI sy'n defnyddio data personol yn risg sylweddol i amrywiaeth ieithyddol yn y dyfodol agos. Er bod cymhathiad AI yn ein bywydau yn y dyfodol agos i weld bron yn anorfod, mae modd lleihau bygythiad penodol unffurfiaeth ieithyddol. Mae'n bosib lleihau'r bygythiad drwy addasu ein hieithoedd i fod yn gymhathus gyda'r modelau AI.

Rwy'n dadlau dros gydymdrechu ar draws y sectorau cyhoeddus, diwydiannol, gwyddonol a chymdeithasol. Mae'r alwad gyffredinol hon i weithredu yn hanfodol i sicrhau bod ieithoedd dynol yn goroesi, yn hytrach na boddi, yn y llanw AI.

Oes gennych chi ffôn symudol?

Fel aelod o'r Genhedlaeth X, rwy wedi bod yn dyst i ddyfodiad ffonau symudol a ffonau clyfar. Mae'r profiad hwn wedi f'arwain at weld cymariaethau rhwng ymatebion cymdeithas i ffonau symudol 25 mlynedd yn ôl a'r ymateb cyfredol i AI. Mae'n ymddangos i mi fod ymateb cymdeithas i'r chwyldroadau technolegol hyn yn debyg iawn. Yn y 1990au hwyr, roedd yna unigolion a oedd yn amheus iawn o ffonau symudol, ond yn awr maent yn eu cario bob dydd. Rwyf wedi catalogio yn ofalus yr amrywiol ffactorau wnaeth gyfrannu at y newid agwedd hon, gan olrhain ymdreiddiad technoleg i gymdeithas o'i gyfnod cychwynnol hyd at gwblhau'r integreiddiad a'r pwyntiau di-droi-nôl ar hyd y daith.

Gadewch i ni feddwl am y newid yma. Gall y rhai sydd wedi profi'r sifft yma, yn enwedig y rhai oedd yn wrthwynebus i ffonau symudol ar y dechrau, gofio'r union foment y gwnaethon nhw ildio i'r dechnoleg newydd. Yn yr un modd, gall y rhai sy'n arsylwi ar ddatblygiad AI ganfod tebygrwydd gyda'r cyfnod cyfredol.

Ar y dechrau, mae bob amser yn…

1 OES YR YCHYDIG BREINTIEDIG

Nodwedd o'r cyfnod hwn yw bod technoleg yn cael ei mabwysiadu'n gyflym gan grŵp bychan o ddefnyddwyr brwdfrydig sydd â chryn dipyn o rym economaidd, diwylliannol a thechnolegol. Mae mabwysiadu'r dechnoleg fel hyn yn rhoi iddi statws symbolaidd o safle cymdeithasol uchel. Mae 'Oes yr Ychydig Breintiedig' yn parhau os yw'r dechnoleg yn parhau i fod y tu hwnt i gyrraedd y mwyafrif. Ar y llaw arall, os daw'r dechnoleg yn fforddiadwy i lawer, yna caiff ei disgrifio fel technoleg sy'n 'democrateiddio'. Mae'r term hwn yn gamarweiniol gan nad oes iddi unrhyw werth democrataidd; mae'n dynodi yn unig nad yw'r cynnyrch mwyach yn gyfyngedig i grŵp elît. Nid yw'r broes hon yn ymwneud â threfn ddemocrataidd fel etholiadau cenedlaethol neu refferenda i benderfynu mabwysiadu technoleg newydd. Grymoedd y farchnad sy'n gyrru'r newidiadau hyn yn bennaf. Roedd y cyfnod hwn yn ddiamau yn amlwg yn achos ffonau symudol, ac ers dyfodiad ChatGPT, mae wedi bod yn fwyfwy gwir am AI. Sylwer fod ChatGPT yn cynnig rhai o'i wasanaethau am ddim, yn wahanol i ffonau symudol, gan gyflymu'r ffordd y mae'n lledu yn fwy eang.

2 GOFYNION PROFFESIYNOL

Mae technoleg yn ysgafnhau nifer o dasgau proffesiynol ac wedi dod yn rhan annatod o nifer o swyddi. Mae holl-bresenoldeb ffonau symudol wedi helpu pobl i fod ar gael bob amser ar gyfer tasgau'n ymwneud â gwaith, datblygiad gafodd ei groesawu ar y dechrau ond a ddaeth wedyn i fod yn anghenraid mewn nifer o broffesiynau, yn enwedig mewn meysydd fel gofal iechyd lle mae'n rhaid i feddygon sydd ar alwad fod ar gael yn barhaus. Yn enwedig mewn oes lle ceir diweithdra eang, gall gwrthsefyll mabwysiadau technoleg ffonau symudol olygu gorfod gadael y proffesiynau hyn. Mae AI yn gynyddol yn galluogi rhai cwmnïau i gymryd y blaen mewn nifer o sectorau, gan gynnwys rhaglennu, creu cynnwys, a gofal cwsmer. Mae effeithiolrwydd AI ynddo'i hun hefyd yn bygwth cynnydd posibl mewn diweithdra, gan ddarparu'r modd a'r cyd-destun ar gyfer cyflymu'r effaith hon.

3 LLEIHAU DIFFYGION SYLWEDDOL

Yn y dyddiau cynnar, roedd yna ganfyddiad cyffredin fod ffonau symudol yn bethau 'peryglus ac aneffeithiol'. Ar y dechrau roedd eu symudedd yn destun sbort, am fod cylch derbyn signal yn gyfyng iawn, yn bennaf mewn ardaloedd bach yn y prif ganolfannau trefol. Ar waethaf pryderon parhaus a sylweddol am beryglon iechyd yn sgil ymbelydredd o'r modelau cynnar hyn, cynyddodd nifer y defnyddwyr a'r ardaloedd sy'n derbyn signal yn sylweddol. Fodd bynnag rhaid nodi fod dau brif bryder o gyfnod cychwynnol ffonau symudol dal yno:

1. mae unigolion sy'n sensitif i donnau electromagnetig yn dal i geisio lloches yn yr ychydig 'barthau gwyn' sy'n parhau, yn aml heb fawr o gonsyrn gan gymdeithas am eu lles, a
2. mae derbyniad signal ffonau symudol yn dal yn anwastad mewn rhai llefydd. Ar waethaf yr heriau parhaus hyn, caiff ffonau symudol eu defnyddio'n eang. Maent yn 'ddigon da'.

Yn yr un modd, yn 2024, mae technoleg AI, sy'n cynnwys llwyfannau fel ChatGPT, weithiau yn cynhyrchu atebion cyfeiliornus, ac nid oes modd gwybod i sicrwydd eto a fydd modd dileu'r 'drychiolaethau' hyn yn llwyr. Mae datblygwyr AI yn gwybod nad yw cyflwyno technoleg berffaith yn angenrheidiol iddi gael ei mabwysiadu yn eang; yn hytrach, y nod yw cyrraedd lefel o weithredu sy'n cyrraedd y trothwy o 'ddigon da' ar gyfer y rhan fwyaf o ddefnyddwyr. Mae hi eisoes yn 'ddigon da' i filiynau o

ddefnyddwyr fwydo rhyngwynebau fel chatGPT gyda llwythi o ddata personol a diwydiannol ac adborth dynol ar y canlyniadau, i'r dechnoleg wella, ac i gyfalaf lifo i mewn i leihau gwallau.

4 OFNAU AC ARGYFYNGAU

Mae ofn yn gatalydd grymus ar gyfer newid mewn cymdeithas. Gall y defnydd o systemau cyfathrebu mewn argyfwng, a hygyrchedd gwybodaeth hanfodol wrth grwydro oresgyn tueddiadau ystyfnig i beidio newid. Mewn sefyllfaoedd argyfyngus, mae'r manteision posib, megis cael at help neu wybodaeth hanfodol heb oedi, yn aml yn drech nag ystyriaethau eraill. Mae hyn yn arbennig o wir lle mae eich diogelwch eich hun neu bobl agos atoch yn y fantol, yn enwedig y mwyaf bregus yn ein plith.

Mae'n amlwg y bydd rhaglenni AI yn cael eu hystyried fel pethau sydd er lles ein hiechyd a'n diogelwch, yn debyg i'r canfyddiad o watsys clyfar sy'n monitro arwyddion iechyd. Mae hyn yn arbennig o wir am blant, henoed ac unigolion sydd ag anableddau. Mae'r poblogaethau hyn yn aml ar y blaen mewn treialon integreiddio technolegau teclynnau agos at y corff, megis mewnblaniadau. Gall y treialon hyn herio tabŵs cymdeithasol dwfn mewn ffordd bragmataidd lle mae integreiddiad corfforol technolegau yn y fantol.

5 GWELLA CYSWLLT DYNOL TRWY DECHNOLEG

Mae bodau dynol wrth eu bodd yn siarad â'i gilydd, ac mae hyn wedi rhoi cyfle i dechnolegau newydd ddod i'n bywydau. Mae ffonau clyfar, er enghraifft, wedi galluogi teulu a ffrindiau i gysylltu'n sydyn ac yn rhwydd drwy negesu a rhyngweithio cymdeithasol. Gyda'r cynnydd yma cafwyd cysylltiad gwell a chynt rhwng pobl a'i gilydd, yn rhagori llawer ar yr hyn roedd yr hen ffonau traddodiadol dros y gwifrau yn ei gynnig. Mae ffenomen Ofn Colli Allan neu FOMO (Fear of Missing Out) wedi dwysáu'r duedd hon ymhellach. Ymhlith cenedlaethau iau, gwelir hyn yn y pryder am golli digwyddiadau cymdeithasol neu ddiweddariadau o fewn eu cylchoedd cymdeithasol. I'w rhieni, roedd yn aml yn adlewyrchu pryder am golli cysylltiad gyda'u plant. Yn yr oes cyn mabwysiadu'r rhyngrwyd yn eang, mewn sefyllfaoedd lle roedd llawer o ddiweithdra, i oedolion, y gallu i glywed am gyfle am swyddi drwy alwad ffôn oedd bwysicaf, cyn twf apiau oedd yn galluogi pobl i gael dêt neu gwrdd â chymar newydd.

O ran AI, yn hytrach na'n pellhau oddi wrth y bobl sy'n annwyl i ni, mae i fod i gynnig dull gwell o gysylltu â phobl. Mae sgwrsfotiaid yn cynhyrchu iaith debyg i iaith ddynol ar unrhyw adeg, heb flino a heb unrhyw gyfyngiad. Maent yn medru cynhyrchu iaith fel bodau dynol go iawn, ond heb ddiffygio a methu fel mae pobl yn gwneud.

Apêl y technolegau hyn yw eu haddewid o fwy o ryngweithio rhwng pobl â'i gilydd. Y nodwedd hon hefyd sy'n eu galluogi i gasglu data personol helaeth, gan ein bod yn aml yn defnyddio'r technolegau hyn yn ein perthynas fwyaf preifat gydag eraill. Mae'n amlwg fod dinasyddion yn dod yn fwyfwy hyddysg yn y gwaith o ddiogelu eu data personol. Mae yna rai sy'n dadlau fod cynnydd yn integreiddiad y dechnoleg ar yr un pryd yn hybu gwrthwynebiad cymdeithasol tuag at y technolegau hyn. Fodd bynnag, rwy'n credu fod y persbectif hwn yn drysu rhwng gwrthwynebiad i weithredu'r technolegau hyn gyda'r trawsnewid sy'n digwydd i'n diwylliannau dynol mewn ymateb i'r technolegau hyn. Yn hytrach na dangos ei hun fel gwrthwynebiad, mae'r addasu diwylliannol hwn yn hwyluso eu hintegreiddiad hapus. Yn ei hanfod mae'n gosod sylfeini ar gyfer hyfywedd rhaglenni sy'n dibynnu ar ddata personol.

Yn y cyfnod nesaf, gan ddechrau gyda phwynt 6, byddwn am gynnig ein data mwyaf personol ac agos atom iddo. Dyma lle bydd siaradwyr ieithoedd sydd ddim yn gydnaws gyda AI yn dechrau cael eu cosbi.

6 HWYLUSTOD AC EFFEITHLONRWYDD

Mae'r gosodiad, "Dw i ddim yn hoffi'r pethau hyn, ond rwy'n hoffi gwaith papur hyd yn oed yn llai," yn adlewyrchu teimlad sy'n cael ei rannu gan lawer. Mae llawer yn cyfaddef fod yr awydd am hwylustod wedi gyrru'r newid tuag at ffonau symudol, gan fod y dyfeisiau hyn yn galluogi cario allan dasgau cyffredin bob dydd. Yn wir, mae ffonau clyfar yn cynnig dewis helaeth o raglenni a gwasanaethau sy'n hwyluso bywyd bob dydd. Ymhlith y rhain mae defnyddio GPS i ganfod y ffordd, rheoli amser drwy galendrau integredig, a thrafod a rheoli arian drwy fancio ar y ffôn.

Mae rhaglenni AI, a gynlluniwyd i ddefnyddio data personol, yn anelu at fodloni'r galw am reoli nifer o dasgau undonog mewn modd cydlynol. Yn y sector gofal iechyd, mae cael at ddata personol yn uniongyrchol yn galluogi integreiddio gwahanol baramedrau iechyd a'u dadansoddi mewn amser real. Mae'r rhain yn cynnwys metrigau corfforol, arferion dietegol, lefelau straen (wedi'u meintioli a'u dadansoddi), cyfansoddiad genetig, a hyd yn oed amodau amgylcheddol megis y tywydd. Mae hyn yn hwyluso diagnoses cychwynnol, ac mae modd i systemau AI staff meddygol proffesiynol eu harchwilio ymhellach. Pwy na fyddai eisiau gallu manteisio ar y buddion hyn?

7 CANOLI TECLYNNAU

Wrth integreiddio gwahanol wasanaethau a ffwythiannau i ffonau symudol, gan gynnwys clociau larwm, camerâu, e-bost, negesu, cyfryngau cymdeithasol, a hyd yn oed tablau'r trai a'r llanw, crëwyd canfyddiad o ddod yn rhydd oddi wrth amlder dyfeisiau unigol, gan y gall un ffôn symudol eu disodli i gyd. Mae rhaglenni AI, gan ddefnyddio data personol, yn ceisio personoleiddio profiadau drwy addasu at ddiddordebau, ymddygiad a dewisiadau unigolion, a thrwy hynny reoli peth wmbredd o apiau mewn dull sy'n addas i'ch arddull a'ch rhythm chi eich hun. Yn y pen-draw, mae hyn yn anelu at ryddhau defnyddwyr o'r holl apiau niferus ar eu dyfeisiau symudol. Mae'n ymddangos yn annhebygol y byddwn yn gallu ymwrthod â'r duedd hon.

8 MYNEDIAD AT WYBODAETH

Drwy integreiddio'r rhyngrwyd gyda ffonau clyfar cafwyd storfa anferthol o wybodaeth sydd ar gael yn hawdd, ac mae hon yn ffactor ysgogol sylweddol. Mae'n debyg fod unigolion a oedd ar y dechrau yn amheus ynghylch ffonau symudol wedi mesur eu pryderon ochr yn ochr â manteision cael mynediad uniongyrchol at newyddion, gwybodaeth a gwahanol adnoddau.

Mae corfforaethau mawr a llywodraethau yn gweithredu mecanweithiau hidlo gwybodaeth, gan ddefnyddio algorithmau i rag-ddewis y wybodaeth sydd ar gael drwy wahanol sianeli, megis yr hidlau presennol sy'n cael eu defnyddio gan Facebook. O gael data personol, mae gan AI y gallu i chwyldroi'r broses hidlo hon. Mae AI yn addo rheolaeth unigol well, gan alluogi defnyddwyr i deilwra eu profiad rhyngrwyd ar sail ffynonellau maent yn ymddiried ynddynt a diddordebau personol, yn lle dibynnu ar hidlau a ddiffiniwyd ymlaen llaw gan eraill, gan gynnwys gwleidyddion a hysbysebwyr. Sylwch y bydd hidlo yn nes i fyny'r ffrwd mwy na thebyg yn parhau, ond bydd yr unigolion yn medru gwrthod mwy. Er enghraifft, bydd unigolion sydd â ffobia o nadredd yn medru osgoi dod ar draws delweddau ohonynt ar-lein. Bydd y rhai sy'n poeni am dabŵs diwylliannol nad yw cymdeithas yn eu cydnabod yn eang yn medru gwe-lywio yn fwy cyfforddus. Bydd eu cynorthwyydd digidol yn cau allan y fath gynnwys ar alw. Felly bydd mynediad pob unigolyn at fyd gwybodaeth yn medru cael ei bersonoli fwyfwy, yn ôl ei ddewisiadau ei hun.

Mae'r cam nesaf yn cychwyn ar bwynt 9. Hwn yw'r cam o golli rheolaeth, gyda phwyntiau di-droi-nôl a chyfyngiadau ar ddewisiadau unigolion.

9 SOFRANIAETH DATA, DIOGELEDD

Yn y blynyddoedd diwethaf, mae ffonau clyfar wedi datblygu i gynnwys technolegau dilysu biometreg ac amgryptio uwch. Yn arbennig, mae gwelliannau mewn gosodiadau preifatrwydd, caniatadau rhaglenni, a rheolaeth defnyddiwr dros fynediad at ddata wedi cael mwy o amlygrwydd. Ar y dechrau, arweiniodd pryderon ynghylch preifatrwydd at betruster ymhlith defnyddwyr; fodd bynnag, mae'r datblygiadau diweddar hyn wedi gwella hyder mewn diogelu gwybodaeth bersonol. Sylwch fod y cynnydd hwn mewn hyder yn parhau er gwaethaf pryderon sydd yno o hyd, megis tracio cyson ar leoliad daearyddol defnyddwyr. Mae unigolion yn teithio yn ddyddiol neu yn rhyngwladol, gan gario gwybodaeth sensitif yn amrywio o fanylion banc i ddata personol iawn ar eu dyfeisiau, sy'n cael eu hystyried i fod yn 'gelloedd digidol' diogel. Dydi colli ffôn clyfar ddim yn beth anghyffredin. Mae gan rai awdurdodau Tollau Cartref a Thramor hefyd yr hawl cyfreithiol i gael mynediad cyflawn at y data hyn. Dyma lle mae ein sefyllfa yn gynyddol debyg i baradeim caethiwed i sylweddau, lle mae'r awydd am dechnoleg yn drech nag ystyriaethau diogeledd critigol. Caiff ein pryderon ynghylch diogeledd yn aml eu tawelu gan sicrwydd arwynebol unwaith fod angen i ni gael mynediad ato (neu unwaith ein bod yn meddwl fod angen i ni gael mynediad). O ganlyniad, rydym yn tueddu i esgeuluso rheolau sylfaenol diogeledd personol neu anwybyddu mannau gwan sylweddol a allai effeithio ar ein rhyddid sylfaenol.

10 INTEGREIDDIAD CYMDEITHASOL

Wrth i ffonau clyfar gael eu mabwysiadu yn eang, mae'r pwysau cymdeithasol i'w defnyddio wedi dwysáu. Mae ffrindiau, teulu, a chydweithwyr yn aml yn dibynnu'n helaeth ar ffonau clyfar i gyfathrebu a chydlynu. Mae hyn yn ei gwneud hi'n anodd i unigolion beidio defnyddio'r dyfeisiau hyn heb deimlo eu bod wedi'u hynysu neu eu cau allan. Ar y cam hwn, daw technoleg yn rhan annatod o gydberthynas pobl â'i gilydd. Mae technoleg nid yn unig yn ymestyn y gydberthynas hon drwy ei hwyluso. Mae'n rhan ohonynt, ac yn chwarae rhan allweddol yn eu ffurfio. Ar ôl y fath newid diwylliannol, mae rhyngweithiadau dynol sydd wedi'u hamddifadu o dechnoleg yn ddiffygiol ac anghyflawn, yn brin o gefnogaeth y fframwaith ddiwylliannol oedd yn bodoli cyn dyfodiad technoleg.

11 HOLLBRESENOLDEB LLWYR GWASANAETHAU

Mae amrediad eang o wasanaethau hanfodol, gan gynnwys trafnidiaeth, siopa, tocynnau a bancio, wedi dod yn fwyfwy integredig gydag apiau ffonau clyfar. Yn dilyn trosglwyddo i lwyfannau digidol, mae darparwyr gwasanaeth yn aml yn ei chael hi'n gostus darparu ar gyfer cwsmeriaid nad ydynt yn cydymffurfio gyda'r normau digidol hyn. O ganlyniad, mae hawliau mynediad cwsmeriaid nad ydynt yn cydymffurfio at wasanaethau yn raddol yn cael eu cwestiynu, gan arwain at ledu'r gwahaniaethau cymdeithasol. Yn y cam hwn, a chan feddwl yn ôl at ein henghraifft gyda rhaglenni iechyd, efallai na fydd meddygon teulu bellach yn derbyn hyfforddiant i drin cleifion nad yw eu data personol wedi cael ei ragbrosesu gan AI.

12 PWYNTIAU DIWYLLIANNOL A GWYBYDDOL DI-DROI-NÔL

Mae arferion cymdeithasol wedi esblygu i ymgorffori'r defnydd o ffonau clyfar. Mae'n gyffredin erbyn hyn, ac yn aml yn cael ei ystyried yn gwrtais, i gadarnhau apwyntiadau ar y ffordd yno, i newid lleoliad y cyfarfod ar y funud olaf yn ôl dewis, neu roi gwybod i eraill am hyd yn oed oedi bychan. Byddai'n amhosibl cynnal yr arferion hyn, sy'n greiddiol i ddefodau cymdeithasol modern, heb y cymorth technolegol hwn. Mae cenedlaethau a fu'n dystion i'r trawsnewidiad hwn yn cofio amser pan oedd hi'n bosib trefnu apwyntiadau fisoedd ymlaen llaw ac yn gwrteisi cyffredin i gyrraedd ar yr amser ac i'r lle dynodedig heb angen unrhyw gadarnhad pellach. Yn ychwanegol, mae ein gallu cyffredin i gofio rhifau ffôn wedi dirywio wrth i ni ddibynnu mwyfwy ar ffonau clyfar ar gyfer y dasg hon. Mae'r newidiadau hyn yn cynrychioli pwyntiau gwybyddol di-droi-nôl. Byddai gwrthsefyll y duedd hon mewn modd trefnus yn golygu bod angen cydlynu heb offer o'r fath, sgil sydd yn gynyddol yn mynd ar goll. Yr

unig ddewis i unigolion sy'n gwrthsefyll y ddibyniaeth hon yw gwrthod y dechnoleg, er y byddai hynny yn golygu arunigo cymdeithasol.

CASGLIADAU

Nid proffwydoliaeth yw'r ysgrif hon ond yn hytrach darpar ddadansoddiad. Yn y naratif hwn o'n hintegreiddiad i'r chwyldro technoleg symudol, oes yna wahaniaeth sylfaenol mewn gwirionedd, naill ai o ran cyd-destun neu dechnoleg, sy'n ei wahaniaethu o AI? O'm persbectif i ar ddechrau 2024, rwy'n credu ein bod wedi symud gan mwyaf heibio cam 3. Mae diwydiannau mawr yn amlwg yn paratoi ar gyfer datblygiadau pellach, a dydw i ddim yn gweld unrhyw rwystrau yn ffordd yr esblygiad cyfredol hwn o integreiddiad technolegol, sy'n f'atgoffa o'r chwyldro technolegol arall hwn rwyf wedi bod yn dyst iddo yn fy mywyd fel oedolyn. Os, fel fi, rydych chi'n methu gweld fawr o wahaniaeth rhwng mabwysiad y ddwy dechnoleg hon gan gymdeithas – ffonau symudol ac AI – mae i weld yn rhesymegol casglu ei bod hi'n annhebygol y bydd gwrthwynebiad i AI. Bydd dylanwad y don hon yn ymestyn yn gyflym i'n sfferau mwyaf personol, fydd yn debyg yn effeithio hyd yn oed ar ein cyrff ffisegol a'n prosesau gwybyddol. Bydd unrhyw wrthwynebiad unigol neu ar y cyd yn ddim mwy nag addasiadau diwylliannol i realiti newydd. Mae'r esblygiad hwn yn cynnig heriau lluosog, a bydd yn parhau i wneud hynny, gyda'r gymuned wyddonol ar hyn o bryd heb fawr o adnoddau i'w wynebu [1]. Mae technolegau AI yn mynnu galluoedd sylweddol o ran caledwedd sy'n medru prosesu meintiau anferthol o ddata. Mae nifer cyfyngedig o gwmnïau sy'n cystadlu â'i gilydd yn gwneud penderfyniadau sylweddol sy'n effeithio ar ein bywydau a'n dyfodol fel rhywogaeth. Mae'r penderfyniadau hyn yn cael eu gwneud heb fewnbwn democrataidd a heb yr adnoddau mewnol angenrheidiol i ystyried yn llawn oblygiadau cymdeithasol eang eu gweithredoedd. Nid yw'r cwmnïau hyn yn cyflogi gwyddonwyr o'r Dyniaethau i ddechrau delio gyda'r problemau hyn.

Fel ieithydd, mae'r perygl o leihad difrifol mewn amrywiaeth ieithyddol yn arbennig o drawiadol i mi. I ddarlunio hyn gydag un achos penodol, mae gan wladwriaeth Ffrainc un iaith swyddogol, Ffrangeg, a fydd yn sicr yn cael ei hymgorffori i dechnolegau AI. Nid yw'r dros gant o ieithoedd answyddogol, diamddiffyn eraill sydd gan ddinasyddion gwladwriaeth Ffrainc wedi'u harfogi o gwbl i basio heibio rhwystr AI. Mae'r rhan fwyaf ohonynt mewn cyflwr o ansicrwydd digidol, yr ydym yn dal yn ymdrechu i'w fapio [2].

Mae cred eang yn ein cymdeithasau, ar wahân i ychydig gylchoedd dylanwadol ond ynysig, y bydd gwrthwynebiad i ddeallusrwydd artiffisial yn gryf mewn cymdeithas, a bod rhagolygon o chwyldro technolegol yn anghywir a heb wneud cyfrif digonol am y gwrthwynebiad y bydd yn ei gynhyrchu. Yn Llydaw, deuthum ar draws y gred hon mewn cenedlaethau gwahanol iawn ac mewn cylchoedd cymdeithasol tra wahanol, o'r Monts d'Arrée i gylchoedd academaidd technoleg uchel y dinasoedd mawr, yn ogystal ag yn …. y cylchoedd technoleg uchel yn y Monts d'Arrée. Rwy'n parchu'r gred hon am yr hyn ydyw. Fel cred. Rwy'n parchu'r emosiynau y mae'n eu hamddiffyn, a'r bobl sy'n eu teimlo. Ond rwy'n meddwl ei fod yn wrthrychol anghywir. Rwy hefyd yn meddwl fod y gred hon yn beryglus pan mae'n arafu'r mesurau y gallwn eu cymryd o hyd i addasu i'r newid hwn sydd ar ddod. Yn wrthrychol mae'r newid hwn yn anochel. Mae angen i'r siaradwyr fod yn ymwybodol o'r peryglon arbennig sydd gan ddatblygiad AI i'w harferion ieithyddol os na all yr offer newydd weithio yn eu hieithoedd hwy. Mae gwadu hynny yn arafu gweithredu mesurau argyfwng i wrthweithio'r peryglon hyn. Os nad oes modd prosesu ein hieithoedd yn awtomatig, fel ag i integreiddio'r offer newydd, byddant yn wynebu gostyngiad hyd yn oed dwysach yn yr ymarfer ohonynt nag a wnaethant yn yr ugeinfed ganrif. Bydd y rhan fwyaf ohonynt yn diflannu.

Dylai unrhyw un sy'n gwybod unrhyw beth am hanes ieithoedd lleiafrifol yn yr ugeinfed ganrif fod yn wyliadwrus iawn yn awr. Dyma pam. Gwyddom fod yr ieithoedd yn diflannu pan fydd rhieni yn credu fod yr ieithoedd hyn yn beryglus i'w plant. Bydd ieithoedd sydd ddim yn gydweddus gyda AI yn dod yn beryglus i'w defnyddwyr, wrth i integreiddiad yr apiau newydd gynyddu mewn cymdeithas. Rydym wedi gweld y bydd rhaglennu AI y byddwn ni – ni ein hunain – eisiau cynnig eu data personol agos iddyn nhw. Y rhain yw pwyntiau 6 i 8 sy'n cael eu darlunio uchod. Os yw ieithoedd bach yn peri i'r rhaglenni hyn fethu gweithio'n iawn, byddwn yn defnyddio ieithoedd sy'n eu galluogi i weithio'n gywir yn eu lle. Mae hyn eisoes yn digwydd yn ymarferol pan fyddwn yn arddweud neges SMS yn Ffrangeg i fanteisio ar ei allu i ysgrifennu'n awtomatig yn hytrach na gorfod ei deipio mewn Llydaweg. Ond mae graddfa'r osgoi yn mynd i fod yn anferthol, heb ddim byd tebyg o'r blaen. Pan fydd cymdeithas yn cyrraedd y pwynt di-droi-nôl, y pwyntiau rhwng 9 a 12 o integreiddiad llwyr na fydd modd ei wyrdroi, bydd y pwysau i gael gwared â ffynonellau nad ydynt yn gweithio'n iawn yn cynyddu. Gyda diffyg cydymffurfio daw allgau cymdeithasol, oherwydd mae'r offer hyn yn treiddio at galon perthnasau rhwng pobl â'i gilydd. Pan fydd y rhaglenni hyn yn dod yn hanfodol i gael mynediad at fancio, cyfiawnder, addysg, cwrdd â chymar newydd a gwasanaethau iechyd, byddwn yn ceisio diogelu dyfodol ein plant drwy eu gwahardd yn anfoddog rhag defnyddio ein hieithoedd sydd ddim yn gydweddus gyda AI, a fydd yn cael eu trosi yn rhwystr cymdeithasol difrifol, mewn symudiad byd-eang mawr tuag at unieithrwydd. Bydd hanes yn dweud ein bod wedi cydsynio â hyn.

Mae yna rhwng 6,000 a 7,000 o ieithoedd yn dal i fodoli yn y byd. Maent yn gynnyrch ein cyrff dynol, biliynau o ymenyddiau dynol yn rhyngweithio â'i gilydd dros filoedd o genedlaethau. Mae diflaniad pob un iaith yn creu clwyfau dwfn, ac yn atseinio dros sawl cenhedlaeth. Dychmygwch gael eich magu gan bobl a wrthododd siarad eu hiaith frodorol eu hun gyda chi. Ac rydym yn wael iawn am wella'r clwyfau hynny. Mae gan ein cenedlaethau ni y cyfle i weithredu i atal y niwed hwn, i sicrhau nad yw'r amrywiaeth hwn yn cael ei ddileu, a bod AI yn gyfle am amlieithrwydd a diogelu'r amrywiaeth anhygoel hwn. Yn sicr mae ganddo'r gallu i wneud ieithoedd yn gydweddus gyda'r dechnoleg hon.

Mae angen cynllun achub byd-eang arnom, symudiad cyhoeddus, gwyddonol a chymdeithasol bywiog, cynllun achub cydweithredol ar gyfer ein trysorau ieithyddol. Yn ymarferol, mae ar ieithoedd angen set data ddigidol o iaith ysgrifenedig a llafar, wedi'i gosod yn y ffurf gywir ar gyfer prosesu digidol [3]. Mae nawr yn hanfodol ac yn fater brys i gymunedau ieithyddol adeiladu eu corpora o iaith lafar ac ysgrifenedig a'u gwneud yn hygyrch ac yn agored, i wyddonwyr a diwydiant fel ei gilydd, gyda hawlfraint agored. Efallai eich bod yn meddwl fod y brys yn tarddu o'u gweithredoedd hwy yn y man cyntaf, ac efallai eich bod chi hefyd yn ddig iawn. Fodd bynnag, gwrthsafiad moesol yn unig fyddai eu hatal rhag cael at y deunydd i integreiddio AI ar gyfer eich ieithoedd chi, ac ni fyddai'n effeithiol yn gwella'r sefyllfa. Ni fydd gwrthsafiad effeithiol i AI. O gymryd ieithoedd yn wystlon gallai hynny selio eu ffawd. Mae nofwyr yn gwybod pan fydd y don yn rhy fawr fod yn rhaid i chi blymio dani, nid dim ond sefyll yno a dweud na ddylai'r don fod yno. Mae angen i bolisïau cefnogi iaith ailfeddwl sut i ddefnyddio'u hadnoddau a chefnogi'r offer achub digidol angenrheidiol [4]. Mae angen iddynt greu cydweithrediad trawsddisgyblaethol rhwng pobl sy'n gwybod sut i godio, addysgu AI mewn ieithoedd sydd heb lawer o gorpora, ac i bobl sy'n gwybod sut i siarad gyda phobl, cynaeafu deunydd ieithyddol mewn dull moesegol, a'i gyfoethogi gydag anodiadau safonol manwl. Mae angen i gymunedau ieithyddol ymrymuso er mwyn adeiladu'r adnoddau digidol sydd eu hangen i gynnwys eu hieithoedd yn yr offer newydd hyn.

O ran y cymunedau gwyddonol sy'n cynhyrchu, prosesu a gweithio ar gorpora, fy nghydweithwyr annwyl iawn, ieithwyr disgrifiadol a ffurfiol, gweithwyr maes, ieithwyr corpws ac ennyn, anodwyr a pheirianwyr pensaernïaeth data, mae gennym gyfle hanesyddol i ddiogelu gwrthrych ein hymchwil ar gyfer cymdeithasau dynol.

Mae'n amser gweithredu a dangos beth allwn ni ei wneud.

DIOLCHIADAU

Hoffwn ddiolch i Rayan Ziane, Loic Grobol, Reun Bideault a Milan Rezac am eu hadolygiadau beirniadol a'u trafodaethau ar y mater. Fi piau unrhyw gamddealltwriaeth neu ddiffygion wrth eu dehongli.

CYFEIRIADAU

[1] Samuel R Bowman. 2023. Eight Things to Know about Large Language Models. Ebr. 02, 2023. arXiv:2304.00612.

[2] Mélanie Jouitteau, Sylvain Kahane a Loic Grobol. 2023-pres. Entrelangues, second edition, IKER & Modyco, CNRS. Adalwyd o https://entrelangues.modyco.fr/.

[3] 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 ACL, Uppsala, Sweden, 88–97.

[4] Mélanie Jouitteau. 2023. Guide de survie des langues minorisées à l'heure de l'intelligence artificielle : Appel aux communautés parlantes. Lapurdum, numéro spécial 6. Adalwyd o https://hal.science/hal-04090195v2.