DOETHUR MEWN ATHRONIAETH

Modelau Cyfrifiadurol ar gyfer Prosesu Lleferydd Cymraeg

Marshall, Indeg

Award date:
2021

Awarding institution:
Bangor University

Modelau Cyfrifiadurol ar gyfer Prosesu Lleferydd Cymraeg

PRIFYSGOL
BANGOR
UNIVERSITY

PhD Ieithyddiaeth

Indeg Marshall

Ysgol Ieithoedd, Llenyddiaethau ac Ieithyddiaeth

Prifysgol Bangor

Tachwedd 2020

Crynodeb

Pwrpas y traethawd hir hwn yw asesu sefyllfa bresennol adnabod lleferydd Cymraeg gan gymharu effaith gwahanol fathau o fodelu, effaith y maint a math o ddata ar gyfer hyfforddi, ac effaith profi ar ddata o acenion gwahanol.

Mae adnabod lleferydd yn rhan o gynorthwywyr personol deallus megis Apple Siri, Amazon Alexa, a Google Home, oll yn dibynnu'n drwm ar gorpora preifat cwmnïoedd masnachol mawr. Oherwydd argaeledd y data, mae ymchwil ym maes adnabod lleferydd yn aml wedi'i seilio ar Saesneg, ar gyfer Saesneg. Mae data addas yn brin ar gyfer y Gymraeg, ac er bod modd edrych i ffynonellau sy'n bodoli'n barod er mwyn casglu mwy o ddata, mae costau a chyfyngiadau trwyddedu yn gallu bod yn rhwystredig. Ystyriaeth bwysig arall yw amrywio o fewn iaith ar gyfer adnabod lleferydd. Mae'r systemau gorau ar y farchnad yn aml yn methu wrth ymdopi ag amrywiadau acen er enghraifft. Wrth ystyried y Gymraeg fel iaith gynhyrchiol, sydd â nifer o ffurfiau gwahanol ac amrywiadau iddi, heb ddata digonol mae hyn yn her ychwanegol ar gyfer adnabod lleferydd.

Nod y gwaith hwn felly yw cymharu sut orau i fynd ati i adeiladu system adnabod lleferydd gadarn, i weld pa faint a math o ddata sydd fwyaf addas, ac i ymchwilio i effaith amrywio iaith o fewn y cyd-destun. Mae'r drafodaeth gyffredinol ynghylch sut orau i fynd ati i gasglu neu i greu data pellach er mwyn cyrraedd y nod hwn.

Cynhelir profion gyda chitiau offer HTK, *Kaldi* a *DeepSpeech*, er mwyn arbrofi gyda modelau gwahanol a deall mwy am sut maent yn ymdopi gyda'r Gymraeg. Defnyddir cyfuniad o ddata hyfforddi geiriau unigol corpws Paldaruo a brawddegau llawn corpws *Common Voice-cy* er mwyn arbrofi gyda mathau a meintiau amrywiol. Hefyd, defnyddir data profi o wahanol acenion (de a gogledd) o raglenni teledu S4C er mwyn astudio effaith amrywio iaith ar system adnabod lleferydd.

Dengys y canlyniadau mai defnyddio modelau ystadegol o fewn cit offer croesryw *Kaldi* yw'r ffordd gorau o gael canlyniadau addawol pan fo data yn brin, gan fod canlyniadau ystadegol HTK a phen-i-ben *DeepSpeech* yn rhoi canlyniadau gwaeth wrth hyfforddi a phrofi gyda'r un data.

O ran hyfforddi gyda gwahanol feintiau a mathau o ddata, mae'r canlyniadau yn gymysg. Er bod y canlyniadau'n cefnogi ymchwil cynt bod mwy o ddata yn well, mae canlyniadau gwell brawddegau llawn *Common Voice-cy* wrth gymharu gyda geiriau

unigol Paldaruo yn unig yn dangos bod math y data hyfforddi yn cael rhywfaint o effaith ar y canlyniadau hefyd.

O ran yr amrywio rhwng data o'r gogledd ac o'r de, gwelir bod y gogledd ychydig yn anoddach i brosesu ond bod dim gwahaniaeth mawr. Gwelir, fodd bynnag, ddirywiad mawr wrth brofi ar ddata acen S4C, o bosib oherwydd bod defnydd siaradwyr o'r iaith yn amrywio o fewn y data profi a bod hyn ar goll yn y data hyfforddi. Dengys profion pellach bod ychwanegu'r amrywiadau hyn i'r data hyfforddi yn gwella canlyniadau.

Awgryma hyn bod angen parhau i ddatblygu gyda chit offer *Kaldi*, bod angen casglu mwy o ddata, a hynny ar ffurf brawddegau llawn, wrth ystyried bod angen rheoli am amrywio iaith rhwng y data hyfforddi a phrofi. Mae'r angen am gynllunio data addas i gipio'r iaith dan sylw yn dangos bod y llwybr i ymchwil pellach yn un ar gyfer cydweithio rhwng datblygwyr ac ieithyddion, er mwyn deall mwy am ffactorau sy'n effeithio'r canlyniadau. Yn ogystal â hyn, gellid ystyried ymchwil pellach i'r dull trosglwyddo dysgu er mwyn gwneud yn fawr o fodelau niwral yn y dyfodol.

Cynnwys

Rhestr Tablau

Rhestr Ffigurau

Rhestr Hafaliadau

Byrfoddau

Byrfoddau	***Teitl Llawn***
C-V	*k-fold cross-validation*
CC0	*Creative Commons*
CER	*Character error rate*
.csv	*Comma seperated values*
cy	*Cymraeg*
CySill	*Cymorth Sillafu*
DA	*Deallusrwydd artiffisial*
DD	*Dysgu dwfn*
(D)NN	*(Deep) neural network*
DP	*Dysgu peirianyddol*
GALLU	*Gwaith Adnabod Lleferydd Uwch*
GC	*Gwaith Cartref*
GMM	*Gaussian Mixture Model*
GPU	*Graphics Processing Units*
HMM	*Hidden Markov Model*
HTK	*Hidden Markov Toolkit*
LTS	*Letter-to-sound rules*
LVCSR	*Large Vocabulary Continuous Speech Recognition*
MFCC	*Mel-frequency cepstral coefficients*
MLP	*Multi-layer perceptron*
NLP	*Natural Language Processor*
OOV	*Out of vocabulary*
.srt	*SubRip*
SVO	*Subject-verb-object*
TIL	*Testun-i-leferydd*
ToS	*Tipyn o Stad*

Byrfoddau	***Teitl Llawn***
VSO	*Verb-subject object*
W.Acc	*Word accuracy rate*
.wav	*Waveform*
WER	*Word error rate*
WISPR	*Welsh and Irish Speech Processing Resources*

Termau

Cymraeg	***Saesneg***
acen	*accent*
adnabod llais	*speaker recognition*
adnabod lleferydd	*speech recognition*
adnabod lleferydd geirfa eang di-dor	*large vocabulary continuous speech recognition*
allan o'r eirfa	*out of vocabulary*
amnewidiad(au)	*substitution(s)*
amrywiant	*variance*
asesiad anghynhenid	*extrinsic evaluation*
asesiad cynhenid	*intrinsic evaluation*
boniwr	*stemmer*
bri	*prestige*
bwydo ymlaen	*feed forward*
cadarn	*robust*
caniataol	*permissive*
chwilio â llais	*voice search*
cit(iau) offer	*toolkit(s)*
clystyru	*clustering*
colled	*loss*
croesryw	*hybrid*
cyfeiliornad safonol	*standard error*
cyflynol	*agglutinative*
cyfnewid cod	*code switching*
cyfnod(au)	*epoch(s)*
cyfradd ddidol	*bit rate*
cyfradd geiriau cywir	*word accuracy rate*
cyfradd geiriau gwallus	*word error rate*
cyfradd llythrennau gwallus	*character error rate*
cymathu	*assimilation*

Cymraeg	***Saesneg***
cynorthwyydd (wywyr) deallus	*smart assistant(s)*
deintiol	*dental*
deusain(seiniau)	*Diphthong(s)*
dilead(au)	*deletion(s)*
dilyniant-i-ddilyniant	*sequence-to-sequence*
dwywefusol	*bilabial*
dysgu dan oruchwyliaeth	*supervised learning*
dysgu heb oruchwyliaeth	*unsupervised learning*
effaith	*impact*
eilededd(au) (sain)	*(sound) alternation(s)*
estyn	*fetch*
fector(au)	*vector(s)*
felar	*velar*
ffonem(au)	*phoneme(s)*
ffrwydrol	*plosive*
gair deffro	*wake word*
geiriadur ynganu	*pronunciation dictionary*
gorfannol	*alveolar*
gorffitio	*overfitting*
graddfa ddysgu	*learning rate*
gwyriad safonol	*standard deviation*
hyfforddi ar y set brofi	*training on the test set*
iaith(ieithoedd) l(l)ai ei(u) hadnoddau	*low-resource language(s)*
iteriad	*iteration*
lema	*lemma*
maint y llwyth	*batch size*
mewnosodiad(au)	*insertion(s)*
model cymysgedd Gausaidd	*Gaussian mixture model*
model Markov cudd	*Hidden Markov model*

Cymraeg	***Saesneg***
monoffon(au)	*monophone(s)*
parsio bwriad	*intent parsing*
Pellter Golygu	*Edit Distance*
Pellter Levenshtein	*Levenshtein Distance*
pen-i-ben	*end-to-end*
perchnogol	*proprietary*
plyg(iadau)	*fold(s)*
prosesydd iaith naturiol	*natural language processor*
pwysau	*weights*
rheolau llythyren-i-sain	*letter-to-sound rules*
rhwydwaith(weithiau) niwral	*neural network(s)*
rhwydwaith(weithiau) niwral dwfn	*deep neural network(s)*
seinegol gytbwys	*phonetically balanced*
set hyfforddi	*training set*
set brofi	*test set*
synthesis lleferydd	*speech synthesis*
tagio rhannau ymadrodd	*part of speech tagging*
technoleg galluogi	*enabling technology*
testun i leferydd	*speech to text*
torfoli	*crowdsourcing*
triffon(au)	*triphone(s)*
trosglwyddo dysgu	*transfer learning*
uned(au) prosesu graffeg	*graphics processing unit(s)*

Diolchiadau

Ariannwyd y traethawd hir hwn gan ysgoloriaeth KESS, Cronfa Gymdeithasol Ewrop a Llywodraeth Cymru felly diolchaf yn gyntaf iddynt am eu nawdd.

Hoffwn ddiolch i bawb yn yr Uned Technolegau Iaith am eu hanogaeth, yn enwedig i Stefano Ghazzali am ei holl gymorth technegol, ac i Aled Tudur ac S4C am eu cefnogaeth. Diolch hefyd i fy nghydweithwyr yn Llywodraeth Cymru, yn y Gwasanaeth Cyfieithu ac yn yr adran Iechyd, am eu hamynedd wrth i mi geisio cydbwyso fy amser.

Diolch yn fawr i fy nhîm goruchwylio, yn enwedig i Dr Sarah Cooper, Yr Athro Delyth Prys, Dewi Bryn Jones a Gareth Morlais am eu hadborth a'u cyngor wrth i mi roi trefn ar fy syniadau. Diolch hefyd i Dr Eirini Sanoudaki, Dr Peredur Webb-Davies, Dr Rhys Jones, Yr Athro Thora Tenbrink a Dr Jeremy Evas am eu hamser.

Yn olaf, diolch i fy nheulu a fy ffrindiau. Diolch i Elaine a Wilf am roi ail gartref i mi. Diolch i fy rhieni yn arbennig, i Mamgu M a Dadcu E, Mamgu A a Dadcu G ac i weddill fy nheulu, am eu diddordeb a'u ffydd. Mae fy niolch pennaf yn mynd i Gareth, am bopeth.

Datganiad

Yr wyf drwy hyn yn datgan mai canlyniad fy ymchwil fy hun yw'r thesis hwn, ac eithrio lle nodir yn wahanol. Caiff ffynonellau eraill eu cydnabod gan droednodiadau yn rhoi cyfeiriadau eglur. Nid yw sylwedd y gwaith hwn wedi cael ei dderbyn o'r blaen ar gyfer unrhyw radd, ac nid yw'n cael ei gyflwyno ar yr un pryd mewn ymgeisiaeth am unrhyw radd oni bai ei fod, fel y cytunwyd gan y Brifysgol, am gymwysterau deuol cymeradwy.

I hereby declare that this thesis is the results of my own investigations, except where otherwise stated. All other sources are acknowledged by bibliographic references. This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree unless, as agreed by the University, for approved dual awards.

I. Cyflwyniad

1. Cyflwyniad

Pwrpas y traethawd hir hwn yw cynnig asesiad ieithyddol-gyfrifiadurol o dechnoleg adnabod lleferydd (*speech recognition*) Cymraeg fel mae'n sefyll, ac awgrymu sut i'w datblygu ymhellach. Cydnabyddir pwysigrwydd technoleg lleferydd ar gyfer ieithoedd lleiafrifol fel y Gymraeg gan Crystal (2000, tt. 141-143), sy'n nodi ei fod yn un o'r camau allweddol i sicrhau goroesiad iaith yn yr oes ddigidol. Caiff y Gymraeg ei hadnabod fel iaith lai ei hadnoddau (*low-resource language*) yng nghyd-destun technoleg lleferydd, sy'n golygu ei bod yn dioddef o ddiffyg adnoddau megis data, cyllid neu arbenigedd (Besacier et al., 2014). Mae cefnogaeth ar gael ac ewyllys da yn gyffredin ar gyfer datblygu adnabod lleferydd ymysg y gymuned Gymraeg (2.3). Fodd bynnag, mae ymchwil yn y maes yn brin ar gyfer yr iaith, felly fe geir yn y traethawd hir hwn gyfraniad newydd i ymchwil adnabod lleferydd Cymraeg a sut i'w gwella. Ceir tri phrif amcan yn y traethawd hir hwn, i edrych sut i fodelu'r Gymraeg yn gyfrifiadurol, i asesu pa faint a math o ddata ieithyddol i'w ddefnyddio, ac i ddeall mwy am effaith amrywio acen (*accent*) ar adnabod lleferydd Cymraeg. Gwneir hyn er mwyn cynnig canllawiau ar gyfer datblygwyr ac ieithyddion Cymraeg ac ieithoedd eraill, i gasglu tystiolaeth all ychwanegu at wybodaeth polisi, ac i arwain at ymchwil pellach yn y maes ieithyddiaeth-gyfrifiadurol.

2. Cyd-destun Ymchwil

Yn y blynyddoedd diwethaf, cyhoeddwyd cynorthwywyr perthnasol deallus (*smart assistants*) gan gwmnïoedd masnachol Apple, Amazon, a Google er enghraifft, sydd wedi trawsnewid y modd yr ydym yn cyfathrebu gyda'n dyfeisiau drwy adnabod lleferydd. Ers cyhoeddiad Apple Siri, Amazon Alexa, a Google Home er enghraifft, mae modd holi eich ffôn symudol, eich cyfrifiadur, rhannau o'ch car, a'ch cartref, am wybodaeth neu dasgau di-ben-draw. Mae'n arf pwerus wrth ystyried rhai heriau meddygol, wrth drin a monitro cleifion sy'n byw â chlefydau dementia (Mulvenna, et al., 2017) a Parkinson (Jilbab et al., 2019) er enghraifft, ac o fewn cyd-destun addysg, wrth hybu hyder wrth gyfathrebu mewn iaith newydd (Carrier, 2017, tt. 53-56). Fodd bynnag, mae'r rhan fwyaf o'r datblygiadau hyn wedi'u seilio ar Saesneg, ar gyfer Saesneg. Gwelir bod Saesneg, a llond llaw o ieithoedd mwy fel Ffrangeg, yn cael eu cefnogi'n gryf gan

Apple Siri er enghraifft, gyda dogfennaeth cymorth Apple, yn debyg i Amazon a Google, yn pwysleisio gallu Siri i siarad mewn lleisiau ac acenion gwahanol mewn Saesneg Americanaidd, Saesneg Prydeinig, neu Awstralia (Apple, 2020).

Heb adnoddau digonol, mae ieithoedd llai ei hadnoddau fel y Gymraeg yn cael eu gadael ar eu hôl, sy'n gallu arwain at siaradwyr dwyieithog yn troi at Saesneg i wneud yn fawr o'r dechnoleg newydd. Petai cyfathrebu drwy'r Gymraeg yn dirywio yn y modd hwn, gellid colli'r cyfle i drosglwyddo rhan annatod o ddiwylliant Cymreig, symbol o hunaniaeth (Evas, 2013, t. 1), i'r genhedlaeth nesaf. Ceir yn y traethawd hir hwn felly, awgrymiadau a chanllawiau am y ffordd orau i barhau i ddatblygu'r dechnoleg ar gyfer y Gymraeg, iaith leiafrifol (2.1) a llai ei hadnoddau (2.2).

2.1. 'Iaith Leiafrifol', 'Iaith wedi'i Pheryglu'

Fel nodir gan Ethnologue (Eberhard et al., 2020), mae'r label 'lleiafrifol' yn cyfeirio at sbectrwm o sefyllfaoedd, lle mae rhai ieithoedd lleiafrifol, fel y Gymraeg, sy'n cael eu defnyddio gan siaradwyr ond yng nghysgod iaith fwyafrifol, neu rhai eraill sydd ar fin diflannu'n gyfan gwbl. Mae gan yr iaith Gymraeg statws swyddogol o fewn Cymru, sy'n meddwl na ddylid ei thrin yn llai ffafriol na Saesneg (Comisiynydd y Gymraeg, 2011). Mae'r Gymraeg yn cael ei siarad gan tua 562,000 o bobl yn ôl yr *Office for National Statistics* (2013) ac yn cael ei disgrifio fel iaith leiafrifol neu wedi'i pheryglu yn aml, fel noda Prys et al. (2020). Yng nghyd-destun technoleg iaith, mae'r label 'lleiafrifol' yn gallu meddwl bod llai o ddata ar gael yn yr iaith. Mae hyn yn bwydo mewn i'r rhan nesaf, 2.2, lle nad oes digon o ddata er mwyn datblygu technoleg iaith, fel adnabod lleferydd er enghraifft. Lle bod perygl o iaith fwyafrifol yn disodli'r iaith leiafrifol oherwydd diffyg argaeledd, mae posibilrwydd o gael ei cholli wrth i dechnoleg iaith ddatblygu. Mae hwn yn her sy'n wynebu ieithoedd eraill hefyd, fel gwelir yn ffigurau Ethnologue (Eberhard et al., 2020). Amcangyfrifwyd bod 7,111 iaith yn cael eu defnyddio dros y byd, gyda (amcangyfrif o) 348 iaith wedi marw'n ddiweddar. Wrth i niferoedd y siaradwyr leihau neu nifer y sefyllfaoedd addas lle gellid siarad yr iaith yn lleihau, mae risg y bydd yr iaith yn datblygu i fod yn iaith 'wedi'i pheryglu' neu'n marw yn gyfan gwbl. Mae Crystal (2000, t. 11) yn diffinio iaith sydd wedi marw, nid iaith fyw hynny yw, fel iaith sydd heb siaradwyr brodorol. Dros amser mae mwy a mwy o bobl yn cyfnewid eu hieithoedd lleiafrifol am ieithoedd mwy pwerus neu swyddogol, agwedd sy'n cael ei ddylanwadu'n gryf gan globaleiddio, ac fel nodwyd uchod, sy'n cynnwys datblygiadau technolegol y

cyfrifiadur a dyfeisiau cysylltiedig (Austin & Sallabank, 2014). Gall y newid hwn arwain at golled araf o'r iaith, ac yn y pendraw at iaith sydd wedi marw.

2.2. 'Iaith Lai ei Hadnoddau'

Caiff iaith lai ei hadnoddau ei diffinio gan Besacier et al. (2014, t. 87) fel iaith lle mae diffyg rhai agweddau (os nad pob un) yn berthnasol: diffyg system ysgrifennu neu orgraff sefydlog, diffyg presenoldeb digidol, diffyg arbenigedd ieithyddol, neu ddiffyg adnoddau digidol ar gyfer prosesu iaith a lleferydd, megis corpora, geiriaduron digidol, data lleferydd addas, geiriaduron ynganu ac yn y blaen. Cafodd y term ei fathu gan Karuwer (2003), lle cyflwynwyd y diffiniad uchod yn wreiddiol. Dylid pwysleisio fodd bynnag, nad yw pob iaith lai ei hadnoddau yn lleiafrifol, er bod y gwrthwyneb fel arfer yn wir. Mae Besacier et al. (2014, t. 87) yn nodi bod rhai ieithoedd llai eu hadnoddau yn cael eu hadnabod fel ieithoedd swyddogol eu gwlad ac yn cael eu siarad gan nifer fawr o'r boblogaeth, ond eu bod yn dioddef o ddiffyg adnoddau yn un o'r ffyrdd uchod, sy'n arafu cynnydd eu technoleg iaith.

Gellid dadlau bod y label 'llai ei hadnoddau', yn hytrach na 'lleiafrifol' er enghraifft, yn gwell adlewyrchu'r dimensiwn pwysig o ddiffyg adnoddau i greu technolegau iaith yn y cyd-destun hwn. Nododd Besacier et al. (2014, t. 86), yn 2014, bod mwy na 6,900 iaith yn y byd yn dioddef o ddiffyg adnoddau ar gyfer technoleg iaith. Erbyn 2020, mae Amazon Alexa yn cefnogi 8 iaith, Google Home yn cefnogi 13 iaith, ac Apple Siri yn cefnogi 21 iaith, y tri â chefnogaeth ar gyfer ffactorau ychwanegol Saesneg megis acenion Prydain, America neu Awstralia er enghraifft (Templeton, 2020). Dyma syniad felly o'r anghydbwysedd adnoddau sydd, mewn rhai achosion, wedi cymell ieithoedd llai eu hadnoddau i ddatblygu'r adnoddau coll hyn ar gyfer technoleg iaith eu hunain. Gwelir y projectau perthnasol ar gyfer y Gymraeg ym mhennod **III** Cefndir: Technoleg Lleferydd Cymraeg. Ceir enghreifftiau o ieithoedd eraill sy'n dioddef o ddiffyg adnoddau am amryw resymau ym mhennod III 1.1 Ieithoedd Llai eu Hadnoddau a Diffyg Data, ac ym mhennod IV 1.1 Ieithoedd Llai eu Hadnoddau ac Amrywio.

2.3. Cefnogaeth: Cynadleddau, Llywodraeth Cymru a META-NET

Erbyn hyn, mae rhai cymunedau llai eu hadnoddau fel y Gymraeg wedi dechrau datblygu adnoddau eu hunain o ganlyniad i ddiffyg cefnogaeth gan gwmnïoedd masnachol mwy. Yn ogystal â hyn, mae'r gymuned ymchwil academaidd wedi ymuno wrth gynnal cynadleddau rhyngwladol er enghraifft, a chynnal gweithdai er mwyn hybu arloesi. Gweler er enghraifft, *The First Celtic Language Technology Workshop* (Judge et al.,

2014) yn COLING 2014 (*International Conference on Computational Linguistics*), a'r Celtic Language Technology Workshop yn y *Dublin Machine Translation Summit 2019* (Spasić, et al., 2019). Law yn llaw gyda'r cynnydd mewn ymchwil academaidd yn y maes, ceir galw cynyddol am y technolegau hyn gan y cyhoedd, ac ar lefelau lleol, fel gwelir wrth holi ar faes yr Eisteddfod Genedlaethol. Drwy ddylunio a chynnal holiadur, fe holais y cyhoedd am eu defnydd o adnabod lleferydd yn ogystal â gwasanaethau Cymraeg digidol eraill. Cynhaliwyd yr holiadur yn stondin yr Uned Technolegau Iaith, Canolfan Bedwyr ym Mhrifysgol Bangor, ym Mhafiliwn Technoleg a Gwybodaeth Eisteddfod Genedlaethol Cymru, Môn 2017. Roedd lefel uchel o frwdfrydedd ac ewyllys da dros ddatblygiadau technolegol Cymraeg, yn awgrymu bod y cyhoedd a gymrodd rhan yn yr holiadur yn barod i weld datblygiadau pellach yn nhechnoleg iaith ar gyfer y Gymraeg. O'r rheiny gymrodd rhan (41 person), roedd 100% yn barod yn defnyddio gwasanaethau Cymraeg, fel tynnu arian o'r twll yn y wal (65.38% o'r rheiny yn 'aml iawn'), 55.77% yn defnyddio adnabod lleferydd yn gyffredinol (15.38% 'aml iawn'), a 59.61% yn frwdfrydig i gyfrannu eu lleisiau er mwyn cael y gwasanaeth adnabod lleferydd yn Gymraeg (40.38% 'tebygol iawn'). Roedd nifer o sylwadau pellach oedd yn cefnogi'r uchod, er enghraifft, 'Mae cael [technoleg iaith] yn y Gymraeg yn hanfodol i ddyfodol yr iaith' (cyfrannwr rhif 41), ac '[Angen] Cadw technoleg Cymraeg ar gyfartaledd [gyda] Saesneg' (cyfrannwr rhif 31).

Fel rhan o'r ymgais i gynyddu nifer siaradwyr Cymraeg, mae strategaethau Llywodraeth Cymru yn cefnogi datblygiad technoleg iaith Gymraeg. Gwelir hyn yn *Iaith fyw: iaith byw* (2012), *Cynllun gweithredu technoleg Cymraeg* (2017), a *Cymraeg 2050: Miliwn o siaradwyr Cymraeg* (2018). Maent yn tynnu sylw at yr angen am fuddsoddi mewn seilwaith technoleg iaith (2018, tt. 70-71) ac yn pwysleisio bod angen cefnogi ymchwil yn y maes er mwyn gwireddu hyn (2018, tt. 77-78). Mae hwn yn nod a gaiff ei rhannu ar draws Ewrop fel gwelir yn ymdrechion y *Joint Research Centre* yn y Comisiwn Ewropeaidd sydd, ymysg meysydd eraill, yn gweithio ar fonitro a datblygu technoleg iaith (Steinberger, et al., 2014). Maent yn pwysleisio pwysigrwydd cyhoeddi adnoddau technoleg iaith er mwyn hybu busnes, rhannu adnoddau, cynnal amrywiaeth ieithyddol, a diogelu amrywiaeth diwylliannol (Steinberger, et al., 2014, t. 685). Yn ogystal â hyn, mae META-NET, rhwydwaith o 60 canolfan ymchwil o 34 gwlad, yn ceisio creu cymdeithas wybodaeth Ewropeaidd amlieithog. Yn 2013, cyhoeddwyd astudiaeth i dechnolegau iaith oedd yn canolbwyntio ar 31 iaith, *Europe's Languages in the Digital Age* (META, 2013) oedd yn trafod y risgiau a'r cyfleoedd mwyaf yn y maes. Wrth gymharu adnoddau ar

draws yr ieithoedd dan sylw, oedd yn cynnwys Saesneg, Ffrangeg, Basgeg, a'r Gymraeg, gwelwyd bod y mwyafrif helaeth â dim ond ychydig iawn o adnoddau, os nad dim o gwbl. Yn yr adroddiad META-NET ar gyfer y Gymraeg yn benodol (Evas, 2013), nodwyd mai diffyg data ac amrywio iaith yn bennaf oedd yn llesteirio cynnydd technoleg iaith Gymraeg, dwy agwedd a gaiff eu hastudio yn y traethawd hir hwn.

2.4. Diffyg Data, Amrywio Ieithyddol a'r Dulliau Perthnasol

Fel nodir uchod, mae'r rhan fwyaf o ddatblygiadau adnabod lleferydd y byd wedi seilio ar Saesneg, ar gyfer Saesneg, ynghyd â dyrnaid o ieithoedd mawr eraill. Ynghlwm â hyn mae maint mawr o ddata sain; defnyddir 2,000-5,000 awr i hyfforddi systemau mwyaf diweddar Saesneg er enghraifft (manylion isod). I gymharu, dim ond 80 awr o ddata sain addas sydd ar gael ar gyfer y Gymraeg ar hyn o bryd. Mae gwaith Hernandez et al. (2019) er enghraifft, yn awgrymu bod ychwanegu mwy o ddata yn gwella canlyniadau, ond mae gwaith Wu et al. (2007) ar y llaw arall, yn awgrymu bod peth o'r data hyn yn ddiangen. Mae hyn yn awgrymu bod modd ystyried ansawdd cynnwys y data dros ei faint mewn sefyllfaoedd lle bo data yn brin neu'n gyfyng. Daw'r 80 awr o ddata Cymraeg uchod o gydweithio rhwng datblygwyr ac ieithyddion sydd wedi ystyried y cysyniad hwn o ansawdd y data dros ei faint ar gyfer y Gymraeg. Wrth gydweithio gydag ieithyddion, fe gasglwyd corpws ffonolegol gytbwys fel corpws Paldaruo (Cooper et al., 2018) er enghraifft, a datblygwyd dulliau newydd o gasglu'r data hwnnw fel *Common Voice Cymraeg* (-*cy*) (Ardila, et al., 2020). Mae'r traethawd hir hwn yn ymchwilio i ddulliau eraill hefyd er mwyn ymdopi â heriau diffyg data fel k-*fold cross-validation* (C-V) (Jurafsky & Martin, 2019, t. 69) er enghraifft, ac yn trafod posibilrwydd defnyddio trosglwyddo dysgu (*transfer learning*) (Meyer, 2019, t. 148) i wella'r canlyniadau ymhellach yn y dyfodol. Gyda chymorth gan ieithyddion, ystyriaeth bellach ar gyfer y Gymraeg yw sut mae'r geiriau hyn yn amrywio rhwng siaradwyr, er enghraifft rhwng acenion gwahanol. Gwelir yng ngwaith Misnikov (2019) a Huang et al. (2020) bod angen mewnbwn ieithyddion er mwyn gallu ymdopi ag amrywio acen, gan ei fod yn arbennig o heriol i fodelu'n gyfrifiadurol.

3. Amcanion Ymchwil

Pwrpas yr ymchwil a geir yn y traethawd hir hwn yw asesu sefyllfa adnabod lleferydd Cymraeg a chasglu tystiolaeth er mwyn cynnig awgrymiadau am sut orau i ddatblygu a gwella. Ceir tri phrif amcan yn y traethawd hir hwn. Yn gyntaf, edrychir i ba

ffordd o fodelu'r Gymraeg yn gyfrifiadurol sy'n creu'r system adnabod lleferydd mwyaf cadarn (*robust*). Yn ail, edrychir i ddarganfod pa faint a math o ddata ieithyddol sydd orau, ac yn drydydd, i feintioli beth yw effaith acen ar adnabod lleferydd Cymraeg, a sut orau i ymdopi â hyn. Er mwyn cyflawni hyn, cynhelir 4 set o arbrofion (manylion isod). Disgwylir i ganfyddiadau'r traethawd hir hwn weithredu fel canllawiau ar gyfer datblygwyr ac ieithyddion Cymraeg ac ieithoedd eraill, fel tystiolaeth i ychwanegu at wybodaeth polisi, ac fel sbardun fydd yn arwain at ymchwil pellach.

4. Trefn Astudiaeth

Fe geir yn y traethawd hir hwn yr ymchwil gyntaf i edrych ar fodelau, maint a math data Cymraeg, yn ogystal ac ystyried effaith amrywio ar adnabod lleferydd Cymraeg. Er mwyn cyfleu cefndir yr ymchwil, rhennir y trosolwg llenyddiaeth i dair pennod. Ceir trosolwg o gefndir y maes yn gyffredinol ym mhennod II Cefndir: Technoleg Lleferydd, yna manylu ar waith cynt technoleg iaith yn y Gymraeg ym mhennod III Cefndir: Technoleg lleferydd Cymraeg, ac yna manylu eto ar yr iaith Gymraeg o safbwynt ieithyddol ym mhennod IV Cefndir: Amrywio yn y Gymraeg. Gwelir esboniad manylach o amcanion yr ymchwil yn y bennod honno hefyd, gan gynnwys y cwestiynau ymchwil a rhagdybiaethau perthnasol. Er mwyn cyflawni tri phrif amcan y traethawd hir hwn fel a geir uchod, cynhelir pedwar set o arbrofion. Ym mhennod V Citiau Offer cymharir modelau addas, yn ogystal â meintiau gwahanol o ddata. Mae pennod VI Math o Ddata yn edrych ar greu a phrofi mathau gwahanol o ddata. Manylir ar effaith amrywio yn y Gymraeg ar adnabod lleferydd ym mhennod VII Amrywio Iaith, cyn dychwelyd at fodelau ym mhennod VIII *DeepSpeech.* Mae'r bennod honno yn ystyried meintiau a mathau gwahanol o ddata hefyd yn fras. Ceir cyflwyniad i'r her benodol sydd dan sylw ym mhob un o'r penodau hyn, manylion y dull a chanlyniadau'r arbrofion, yn ogystal â thrafodaeth interim. Ceir trafodaeth gyffredinol ym mhennod IX Trafodaeth, lle ystyrir canfyddiadau'r trafodaethau interim, wrth gydnabod unrhyw gyfyngiadau neu ymchwil pellach all arwain ymlaen o'r gwaith hwn.

II. Cefndir: Technoleg Lleferydd

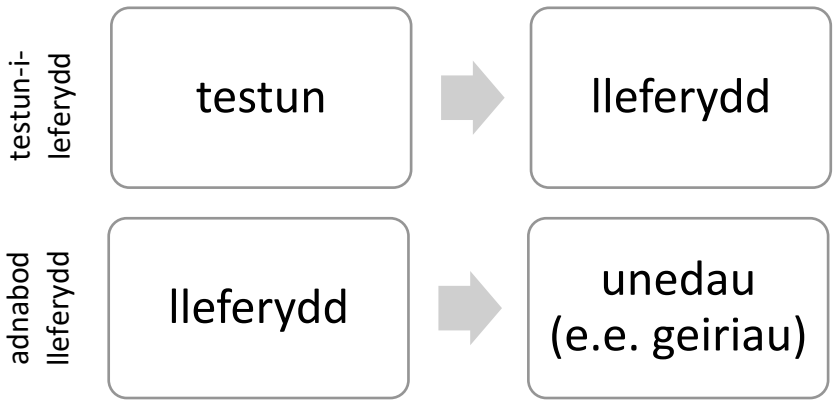
1. Cyflwyniad

Pwrpas y bennod hon yw cynnig cyflwyniad i gefndir y maes technoleg lleferydd. Mae technoleg lleferydd yn cyfeirio at dechnoleg sy'n prosesu lleferydd dynol. Mae'r dechnoleg yn cael ei defnyddio ar gyfer:

- creu llais artiffisial (synthesis lleferydd, *speech synthesis*);
- deall llais person (adnabod lleferydd);
- adnabod llais person (adnabod llais, *speaker recognition/verification*).

Ceir esboniad o'r defnyddiau hyn yn y bennod hon, wrth ganolbwyntio ar adnabod lleferydd, a sut mae modelau gwahanol yn cael eu defnyddio er mwyn eu gweithredu. Mae'r bennod wedi rhannu i dair rhan: adnabod lleferydd yn 2, modelau yn 3, a dulliau i ymdopi gyda llai o ddata yn 4. Gwelir y cwestiynau ymchwil sy'n arwain ymlaen o hyn oll ym mhennod IV 5 Amcanion yr Ymchwil.

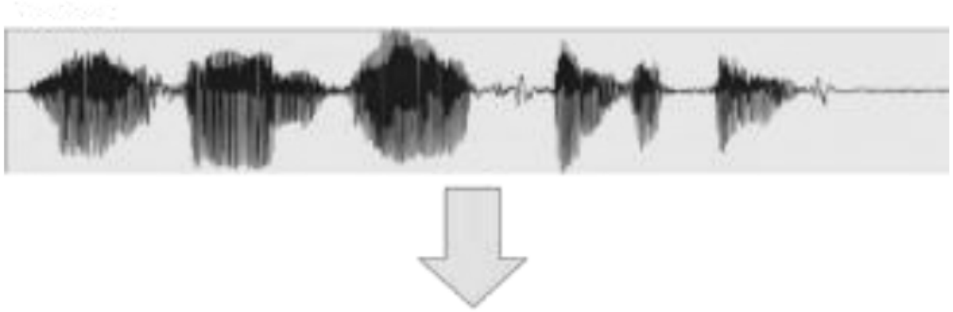
1.1. Cymharu Adnabod Lleferydd a Thestun i leferydd



Ffigwr 1 Gwahaniaeth rhwng testun-i-leferydd ac adnabod lleferydd

Er bod y traethawd hir hwn yn canolbwyntio ar adnabod lleferydd, ceir cymhariaeth rhwng adnabod lleferydd a thestun-i-leferydd (TIL, *speech to text*, neu synthesis lleferydd o'r rhestr uchod) yma gan eu bod, gyda'i gilydd, yn cwmpasu technoleg lleferydd (Jurafsky & Martin, 2019, t. 502). Pwrpas adnabod lleferydd yw trosi mewnbwn lleferydd i linyn o unedau (geiriau gan fwyaf) (Cooper, 2019, t. 1788). Mae TIL fodd bynnag, yn

cysylltu llinynnau o destun i seindonnau acwstig (Jurafsky & Martin, 2019, t. 518), drwy ddefnyddio trawsgrifiadau ffonemig, a'u defnyddio i syntheseiddio lleferydd (Cooper, 2019, t. 1788). Mae adnabod lleferydd a TIL yn rhan o gynorthwywyr deallus megis Apple Siri ac Amazon Alexa er mwyn clywed sain, ei throsi i eiriau, ac yna creu lleferydd yn ôl (Jones et al., 2019, t. 25). Gwelir cynrychiolaeth o donffurf mewn geiriau isod yn Ffigwr 2.



MAE HEN WLAD FY NHADAU

Ffigwr 2 Cynrychiolaeth o donffurf mewn geiriau

(Jones, Prys, Prys, & Prys, 2019, t. 25)

Yn ystod y 1930au, yn Bell Labs yn Unol Daleithiau America, roedd ymchwilwyr wrthi'n ystyried technoleg adnabod lleferydd (Dudley, 1939; Dudley, Riesz, & Watkins, 1939). Roedd cryn dipyn o ansicrwydd ynghylch potensial cywirdeb adnabod lleferydd, gydag un o'r beirniadaethau cyntaf yn ymddangos ym 1969 pan ysgrifennodd Pierce (1969), peiriannydd ac awdur adnabyddus, lythyr cyhoeddus yn beirniadu ymchwil adnabod lleferydd. Fodd bynnag, ym 1971, ailgychwynnodd ymchwil wrth i DARPA (*Defense Advanced Research Projects Agency*) gefnogi pum mlynedd o *Speech Understanding Research* a dynnodd sylw ymchwilwyr o BBN (*Bolt Beranek and Newman Inc.*), IBM (*International Business Machine Corporation*), *Carnegie Mellon* a'r *Stanford Research Insitute* (Klatt, 1987). Arweiniodd y bwrlwm hwn at yr ICASSP (*International Conference on Acoustics, Speech and Signal Processing*) cyntaf ym 1976 (Rabiner, 1984), sy'n dal i fod yn blatfform pwysig ar gyfer cyhoeddi ymchwil newydd yn y maes.

Mae TIL wedi bod yn cael ei ddatblygu ers y 1930au hefyd, yn bennaf gan Bell Labs, gydag iteriad cynnar y *voder* (*Voice Operating Demonstrator*), lle roedd angen pwyso botymau a phedalau i weithredu'r llais, a groniclwyd gan Mills (2012). Fel noda Klatt yn ei drosolwg o TIL Saesneg (1987), datblygodd hwn ymhellach gan ddefnyddio ymchwil Prifysgol Nagoya a'r *Japan Electrotechnical Laboratory* yn y 1960/70au, i feddalwedd DECtalk erbyn yr 1980au a'r 1990au. Pwrpas DECtalk oedd efelychu'r llwybr llais i greu'r llais synthetig a gafodd ei ddefnyddio i helpu'r rheiny oedd ag anghenion arbennig lle nad oedd darllen yn ymarferol (Hallahan, 1995). Mae mwy am dechnoleg TIL Cymraeg ym mhennod III Technoleg Lleferydd Cymraeg.

1.2. Defnyddiau o Dechnoleg Lleferydd

Fel nodir yn 2, defnydd mwyaf poblogaidd y dechnoleg hon yw gallu cyfathrebu â'ch dyfeisiau ac felly, pwrpas yr is-ran hon yw cyflwyno rhai o'r defnyddiau posib lle gallai technoleg lleferydd Cymraeg gael ei defnyddio yn y dyfodol. Ceir enghreifftiau isod o'r meysydd archifo radio a theledu, addysg, ac iechyd.

Archifo Radio a Theledu

Mae gan nifer o sefydliadau megis S4C, Llyfrgell Genedlaethol Cymru a'r BBC, archifau digidol enfawr sy'n cynnwys oriau o recordiadau radio a theledu. Ar y foment, does dim llawer o waith trawsgrifio awtomatig wedi'i gyflawni yn y Gymraeg er mwyn eu catalogio a'u mynegeio er mwyn chwilio. Wrth ddefnyddio technoleg adnabod lleferydd, byddai modd trawsgrifio'r recordiadau sydd yn yr archifau er mwyn eu categoreiddio a galluogi chwilio mynegai am recordiad neu thema benodol. Gwelir canlyniadau addawol (tua 9-10% WER, *word error rate*, cyfradd geiriau gwallus, lle cyfrifir faint o wallau ar lefel geiriau sy'n digwydd wrth hyfforddi a phrofi, V 2.2 Gweithdrefn) wrth ddefnyddio technoleg lleferydd i drawsgrifio ffeiliau BBC Saesneg yng ngwaith Lanchantin et al. (2013), a gwaith Mohr et al. (2013) gyda recordiadau o fwletinau tywydd BBC. Dengys gwaith Jong et al. (2018) hefyd bod modd defnyddio adnabod lleferydd er mwyn echdynnu gwybodaeth (metadata) o archifau lleferydd er mwyn categoreiddio a chwilio, ond yn pwysleisio bod angen ymchwil pellach yn y maes er mwyn gwella'r dechnoleg i safon lle bydd modd buddsoddi yn y dull yn ymarferol.

Addysg

Gellir defnyddio technoleg adnabod lleferydd i helpu dysgwyr ail iaith a phlant ysgol yn eu gwaith dysgu. Fel dysgwr ail iaith, mae'n aml yn anoddach creu sefyllfaoedd i

ddefnyddio'r Gymraeg yn fwy naturiol mewn bywyd dydd-i-ddydd. Dengys gwaith Neri et al. (2003) bod adnabod lleferydd yn fanteisiol, a chost effeithiol (Strik et al., 2008), ar gyfer dysgu ail iaith drwy Saesneg gan ddefnyddio meddalwedd a gweithgareddau dysgu addas, yn enwedig ar gyfer ynganiad. Dengys gwaith Carrier (2017) bod modd defnyddio'r dechnoleg yn y dosbarth ysgol Saesneg hefyd. Gellir ei defnyddio ar gyfer darllen yn uchel, neu dasgau cyfieithu, ac annog hyder wrth gyfathrebu (2017, tt. 53-56). Wrth wneud yn fawr o'r dechnoleg sydd ar gael, gwelir manteision cyflymder a chyfleustra ar gyfer tasgau addysgiadol (2017, t. 59). Atgyfnerthwyd hyn gan Moussalli a Cardoso (2016), gan drafod manteision defnyddio cynorthwywyr deallus yn benodol, gan ddod i'r casgliad eu bod yn ehangu gorwelion dosbarthiadau o ran y deunydd a'r dulliau sy'n cael eu defnyddio. Byddai defnydd technoleg adnabod lleferydd mewn dosbarthiadau Cymraeg yn cefnogi strategaethau mwyaf diweddar Llywodraeth Cymru, fel a gyfeirir atynt ym mhennod I. Mae *Cymraeg 2050: Miliwn o siaradwyr* (2017) yn nodi bod angen 'cynyddu'r gyfran o bob grŵp blwyddyn ysgol sy'n cael eu haddysg drwy gyfrwng y Gymraeg o 22 y cant i 30 y cant', a 'gweddnewid sut rydym yn addysgu Cymraeg i bob dysgwr fel bod o leiaf 70 y cant o'r dysgwyr hynny'n gallu dweud erbyn 2050 eu bod yn gallu siarad Cymraeg' (2017, t. 12).

Iechyd

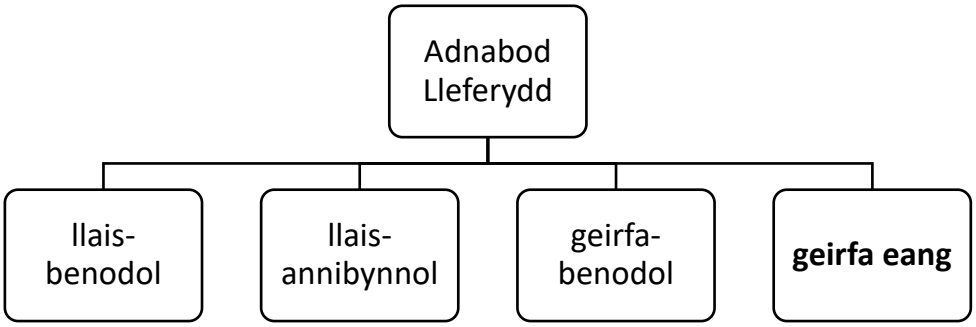
Defnydd posib arall yw o fewn y maes iechyd, i helpu'r rheiny sy'n dioddef o anawsterau gweledol, clywedol, neu symudol. Mae Delić et al. (2014) yn nodi sut gall TIL yn iaith Serbia helpu'r rheiny sydd angen cymorth clywedol tra bod datblygiadau adnabod lleferydd wedi helpu'r rheiny sydd angen cymorth gweledol. Gwneir hyn drwy ddefnyddio technegau arddweud. Mae Garrett et al. (2011) yn nodi sut gall meddalwedd adnabod lleferydd Serbia helpu'r rheiny sydd ag anawsterau wrth gyfleu syniadau ar bapur hefyd. Mae modd teilwra'r dechnoleg ymhellach i helpu cleifion hefyd; gwelir gwaith Mulvenna et al. (2017) gyda chleifion dementia yn y DU (Saesneg), gwaith Jilbab et al. (2019) gyda chleifion Parkinson, a gwaith Costa et al. (2018) ar fonitro cleifion mewn ysbytai (ffocws Saesneg ar hyn oll). Yn ychwanegol i hyn, noda gwaith Zhang & Hansen (2019) bod dulliau technoleg lleferydd o wneud diagnosis lawer rhatach na'r gweithredoedd mewnwthiol sy'n cael eu defnyddio fel arfer. Mae ymchwil ar waith i ehangu hyn i anawsterau dysgu hefyd (Gautam & Singh, 2019), megis gwaith Marchi et al. (2019) a Diehl (2019) gydag awtistiaeth (ffocws Saesneg eto). Un o'r nifer bach o astudiaethau sydd wedi cael eu datblygu ar gyfer ieithoedd eraill ar wahân i'r Saesneg yw

Barbosa et al. (2019) sy'n astudio paramedrau acwstig yn y rhaglen *Praat* (rhaglen i ddadansoddi nodweddion acwstig lleferydd). Mae Barbosa et al. (2019) yn astudio patrymau pwyslais a goslef iaith cleifion er mwyn sefydlu graddfa 'arferol' a gallu gwneud diagnosis o broblem lleferydd. Mae'r adnoddau o'r astudiaeth hon wedi'u rhyddhau yn agored ac felly wedi cael eu profi ar gyfer Ffrangeg, Almaeneg, Swedaidd, Saesneg, a Phortiwgaleg. Mae gwaith ym maes iechyd ar y Gymraeg wedi cael ei wneud gan yr Uned Technolegau Iaith fel rhan o'r project Lleisiwr yn (2018). Mae hwn yn broject sy'n bancio lleisiau cleifion yn synthetig cyn iddynt golli eu lleisiau drwy gyflyrau megis Clefyd Niwronau Echddygol a Chanser y Gwddf. Mae'n broses sy'n digwydd dros y we, lle mae cleifion yn recordio'u llais drwy gyfres o bromptiau er mwyn cael ei ail-greu yn synthetig.

Mae'n bwysig cydnabod bod y rhan helaeth o'r datblygiadau hyn dim ond wedi'u datblygu ar gyfer yr iaith Saesneg. Fel sydd i'w weld ym mhennod I 2.2 'Iaith Lai ei Hadnoddau', mae'r rhan fwyaf o ddatblygiadau'r maes technoleg lleferydd wedi canolbwyntio ar ieithoedd sydd â digonedd o adnoddau, fel Saesneg. Mae hyn yn golygu bod diffyg cefnogaeth ar gyfer ieithoedd llai eu hadnoddau o fewn y maes technoleg 'clyfar'. Pwrpas gweddill y bennod hon yw cynnig cyflwyniad i'r maes technoleg lleferydd, yn enwedig i adnabod lleferydd, er mwyn deall mwy am beth sydd ar gael, ac arwain ymlaen at y bennod nesaf, III Cefndir: Technoleg Lleferydd Cymraeg, sy'n cyflwyno sut mae modd datblygu'r rhain ar gyfer ieithoedd llai eu hadnoddau fel y Gymraeg.

2. Adnabod Lleferydd

Gwelir uchod yn 1, rhai defnyddiau ar gyfer technoleg lleferydd yn gyffredinol, lle rhoddir ffocws arbennig ar adnabod lleferydd. Pwrpas yr is-bennod hon yw cynnig cyflwyniad mwy manwl i'r mathau gwahanol o adnabod lleferydd. Mae modd gwahanu adnabod lleferydd fewn i wahanol fathau fel gwelir yn Ffigwr 3:



Ffigwr 3 Mathau gwahanol o dechnoleg adnabod lleferydd

(wedi addasu o Jones et al. (2019, t. 26))

Mae'r rhan fwyaf o'r mathau gwahanol o adnabod lleferydd uchod yn Ffigwr 3, wedi derbyn sylw ymchwil yn barod (manylion yn III 2.3 'GALLU, Paldaruo a Macsen) ac maent yn gweithredu'n llwyddiannus i wahanol lefelau o effeithiolrwydd. Mae geirfa eang, mewn testun trwm uchod, yn cael ei ystyried fel yr her fwyaf fel mae Jones et al. (2019, t. 26) yn adleisio. Dyma'r math o dechnoleg adnabod lleferydd lle mae angen trosi sain i destun gan unrhyw siaradwr wrth siarad yn ddi-dor, mewn iaith naturiol gydag ystod eang o eirfa, yn hytrach na geiriau cyfyngedig ar bwnc penodol. Ceir manylion pellach am waith ar hyn yn 2.1.

2.1. Hanes LVCSR a Chynorthwywyr Deallus

Gwelir yn Ffigwr 3 bod geirfa eang yn un o'r mathau gwahanol o adnabod lleferydd a gaiff ei gydnabod fel un o'r rhannau mwyaf heriol yn y maes ar gyfer ieithoedd llai eu hadnoddau (Jones et al., 2019, t. 26). Mae math penodol o adnabod lleferydd, geirfa eang di-dor (*large vocabulary continuous speech recognition*, LVCSR) yn golygu datblygu adnabod lleferydd sydd yn gallu ymdopi gydag unrhyw leferydd heb seibiant rhwng geiriau. Dengys ymchwil Jones et al. (1997) pa mor heriol a chymhleth yw'r dasg o greu LVCSR. Gan ddefnyddio *Subscriber*, sef cronfa Saesneg o frawddegau llawn, arbrofwyd gyda modelu fesul sillaf yn hytrach na ffonemau ond gwelwyd gwaethygiad mewn WER (cyfradd geiriau gwallus, V 2.2 Gweithdrefn) o 35% ar y pryd i 60%. Mae'r rhan hon yn edrych sut mae datblygiadau diweddarach yn LVCSR wedi arwain at gynorthwywyr deallus. Fel nodir ym mhennod I 2 Cyd-destun Ymchwil, mae cynorthwywyr deallus wedi sbarduno diddordeb mawr yn y dechnoleg, tu allan i'r gymuned dechnolegol. Gwelir yn 1.2 ei fod yn cynnig manteision yn y meysydd archifo, addysg, ac iechyd, yn

ogystal â chyfleustra cyfathrebu â dyfeisiau personol yn gyffredinol. LVCSR yw'r gydran allweddol i alluogi hyn.

Gan ddefnyddio dulliau ystadegol (manylion isod yn 3), dechreuodd ymchwilwyr gyfuno sgiliau technolegol gyda rhai ieithyddol fel cystrawen a seineg. Gwelir yn hanes bras Juang & Rabiner (2004) bod y fath ddulliau wedi'u defnyddio drwy'r 1980/90au gan gwmnïoedd mawr fel IBM ac AT&T i greu lleisiau electronig megis Dragon Dictate, system oedd yn galluogi llinellau ffôn i gael eu cysylltu heb weithredwr dynol. Ym 1992, dechreuodd ymchwilwyr gasglu'r dulliau hyn a'u cyhoeddi ar ffurf cit offer (*toolkit*) (gwelir pennod V Citiau Offer). Un o'r cyntaf oedd HTK (*Hidden Markov Model Toolkit*) ym Mhrifysgol Caergrawnt (Young et al., 2015). Wrth i ddiddordeb mewn LVCSR gynyddu dros y 1990au, sefydlwyd grwpiau ymchwil adnabod lleferydd ym Microsoft a dechreuodd Apple a Google ymchwilio i ryngwynebau lleferydd. Dyma fyddai'n arwain at y cynorthwywyr deallus caiff eu hadnabod heddiw megis Apple Siri, Amazon Alexa a Google Home. Maent yn ymateb i air deffro (*wake word*) megis 'Hey Siri', 'Alexa' neu 'OK Google', ac mae modd defnyddio ar gyfer tasgau eang fel gofyn am wybodaeth am y tywydd, chwarae cerddoriaeth a rhaglenni, neu ddarllen llyfrau yn uchel. Fel yn 1.2, mae cynorthwywyr deallus hefyd wedi treiddio i feysydd iechyd, yn ogystal â thelathrebu, manwerthu ac addysg, gan wneud yn fawr o hyblygrwydd a rhwyddineb y broses chwilio ac ateb.

Fel nodir ym mhennod I 2 Cyd-destun Ymchwil, mae angen data ar raddfa fawr er mwyn i'r systemau hyn fod yn effeithiol (cyflwyniad ym mhennod I 2.4 Diffyg Data, Amrywio Ieithyddol a'r Dulliau Perthnasol, manylion ym mhennod V 1.1 Yr Her: Diffyg Data). Mae hwn yn heriol i ieithoedd llai eu hadnoddau fel y Gymraeg wrth ddatblygu systemau tebyg, a bydd trafodaeth bellach o'r her ym mhennod III 1.1 Ieithoedd Llai eu Hadnoddau a Diffyg Data.

Deallusrwydd Artiffisial a Dysgu Peirianyddol

O ganlyniad i ddatblygiadau yn y byd symudol fel chwilio â llais (*voice search*), arddweud, a chynorthwywyr deallus, (yn ogystal â data mawr a phŵer cyfrifiadurol) mae deallusrwydd artiffisial (DA) wedi datblygu yn ddiweddar. Mae DA yn cyfeirio at ddeallusrwydd caiff ei arddangos o beiriannau, yn wahanol i'r hyn a geir yn naturiol gan bobl neu anifeiliaid. Mae'r agweddau hyn wedi gweddnewid maes adnabod lleferydd yn lled ddiweddar (Yu & Deng, 2015, t. ix), ac wrth i DA, dysgu peirianyddol (DP), a dysgu

dwfn (DD) wella, mae cynorthwywyr deallus yn 'dysgu' mwy am eu hamgylcheddau ac yn gwella'n gyson.

Gydag anogaeth Turing (Copeland & Proudfoot, 2009, t. 120), datblygwyd y rhaglenni DA cyntaf yn y 1950au, ac fe grëwyd yr efelychiadau cyntaf o rwydweithiau niwral (*neural networks*), yn Labordy Lincoln yn MIT (*Massachusetts Institute of Technology*) (Copeland & Proudfoot, 1996, t. 368). Arweiniodd hyn at dair ton o ddiddordeb ynghylch DA, hyd nes ddiwedd y 1990au, lle enillwyd gêm gwyddbwyll gan y cyfrifiadur *Deep Blue* yn erbyn pencampwr y byd Garry Kasparov (McCorduck, 2004, tt. 480-483). Ffrwydrodd y maes yn ystod y 2010au pan ddechreuodd systemau rhwydweithiau niwral dwfn (*deep neural networks*, DNN) ddechrau perfformio'n well na systemau ystadegol cynt.

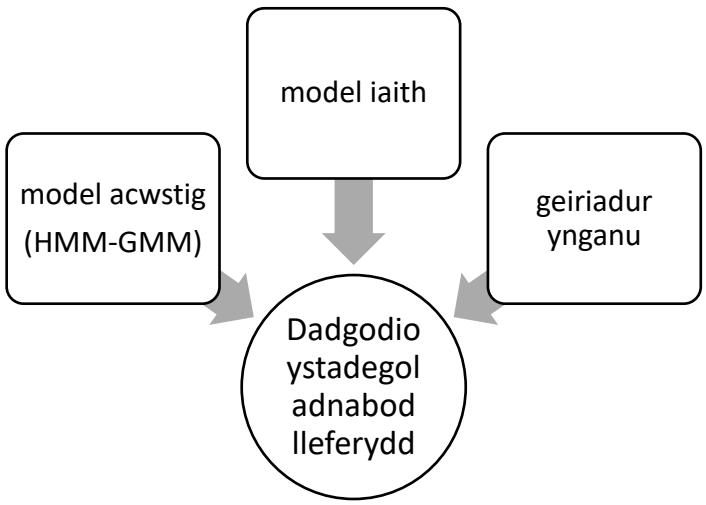
Mae DP yn cyfeirio at beiriant sy'n dysgu o batrymau yn hytrach na chyfeiriadau penodol. Mae'n defnyddio data hyfforddi er mwyn gwneud rhagdybiaethau a phenderfyniadau er mwyn cyflawni tasg. Cafodd y term ei fathu gan IBM oedd yn arloeswr yn y maes gemau cyfrifiadurol a DA (Samuel, 1959). Dan oruchwyliaeth (*supervised learning*, hynny yw, gyda data wedi'i labelu gan y datblygwr), mae'r peiriant yn gweld enghreifftiau o'r mewnbwn a'r allbwn ar ffurf data wedi'i hanodi gyda labeli neu drawsgrifiadau, lle mae angen i'r cyfrifiadur ddarganfod parau a'u hailgynhyrchu (Jones et al., 2019, t. 9). Dim ond y mewnbwn, heb unrhyw anodi neu labeli ar y data, sydd ar gael ar gyfer dysgu heb oruchwyliaeth (*unsupervised learning*), ac mae angen i'r peiriant adnabod patrymau wrth glystyru (*clustering*), er enghraifft adnabod bod 'bryn' a 'mynydd' yn perthyn (Jones et al., 2019, t. 10). Ceir cyflwyniad i'r modelau sy'n cyfateb i'r dulliau hyn isod yn 3.

3. Modelau

Mae'r is-bennod hon yn manylu ymhellach ar y dulliau gwahanol o weithredu adnabod lleferydd o'r fath. Rhennir yr is-bennod i dair rhan gan ddechrau gyda modelau ystadegol yn 3.1, cyflwyno modelau croesryw (*hybrid*) sy'n defnyddio dulliau niwral mwy diweddar yn ogystal â rhai ystadegol yn 3.2, ac yna modelau pen-i-ben sy'n defnyddio dulliau niwral yn unig yn 3.3. Mae dulliau ystadegol yn cyfeirio at ddulliau sy'n seiliedig ar hafaliadau tebygolrwydd (isod), lle mae rhwydweithiau niwral (dwfn) yn efelychu strwythur yr ymennydd yn gyfrifiadurol, ac yn cyfrifo'r tebygolrwydd o grwpio gwybodaeth i grwpiau tebyg a'i phasio trwy haenau nes cyrraedd yr ateb cywir (mwyaf tebygol) (Jones et al., 2019, tt. 7-8).

3.1. Modelau Ystadegol

Er mwyn creu modelau ystadegol, defnyddir modelau Markov cudd (*Hidden Markov model*, HMM), sy'n modelu'r tebygolrwydd o gyfres o arsylwadau neu ffonemau, (*phonemes*). O fewn y cyfresi hyn, defnyddir modelau cymysgedd Gausaidd (*Gaussian mixture model*, GMM) sy'n ddosraniadau o'r unedau hyn (Jones et al., 2019, t. 26). Gelwir y model yn HMM-GMM felly. Maent ill dau yn fodelau ystadegol sy'n cael eu defnyddio ar gyfer y dasg adnabod lleferydd. Mae esboniad manylach o'r broses adnabod lleferydd fesul model ym mhennod V 2.2 Gweithdrefn. Ond i gyflwyno'r cysyniad, defnyddir system adnabod lleferydd gyfuniad o fodel acwstig, model iaith a geiriadur ynganu (*pronunciation dictionary*) i ddadansoddi mewnbwn acwstig fel gwelir isod yn Ffigwr 4.



Ffigwr 4 Proses ddadgodio ystadegol system adnabod lleferydd yn fras

Fel nodir ym mhennod V 2.2 Gweithdrefn, mae model acwstig yn cyfrifo'r tebygolrwydd seiniau, lle mae model iaith yn cyfrifo tebygolrwydd brawddeg gan ddefnyddio cymorth geiriadur ynganu i fapio cyfres o ffonemau i eiriau (enghreifftiau ym mhennod V 2.2 Gweithdrefn) (Cooper, 2019, t. 1789). Defnyddir hafaliad Bayes er mwyn cyfrifo'r tebygolrwydd sef:

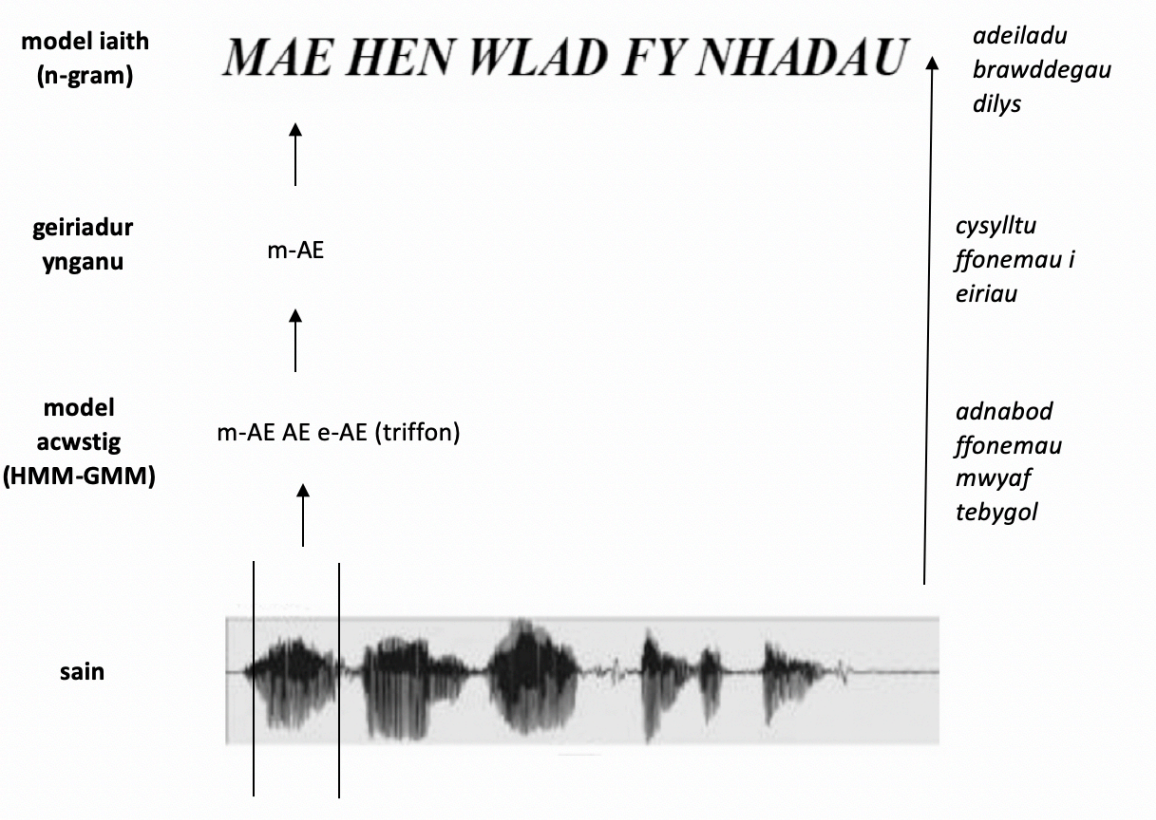
$$P(W|O) = P(O|W)P(W)$$

Hafaliad 1 Hafaliad Bayes

lle ceir y gair (W=*word*) mwyaf tebygol o ganlyniad i'r mewnbwn acwstig (O=*observation*), $P(W|O)$, drwy luosi'r tebygolrwyddau sydd wedi'u cyfrifo o'r data hyfforddi, hynny yw, y model acwstig, $P(O|W)$, a'r tebygolrwydd o'r gair ei hunan, hynny yw, y model iaith, $P(W)$ (Jurafsky & Martin, 2019, tt. 65-66).

Gan fod ffonemau yn aml yn cael eu heffeithio gan seiniau cyfochrog, mae'n gallu bod yn gamarweiniol i ffocysu ar fonoffonau (*monophones*) yn unig, hynny yw seiniau unigol heb ystyried seiniau cyfochrog. I osgoi'r broblem hon, fel gwelir yn Gales & Young (2007, tt. 206-207), defnyddir HMM-GMMs i fodelu ffonemau o fewn cyd-destun triffonau (*triphones*) sy'n edrych i'r ffonemau cyfochrog, er enghraifft "beth": /b/ + /e:/ + /θ/, a dadansoddi tebygolrwydd un ffonem yn y dilyniant. Mae'r cam yma yn angenrheidiol er mwyn galluogi system adnabod lleferydd i ymdopi gydag amrywio (cyflwyniad i eilededdau sain er enghraifft ym mhennod IV 3.3 Tafodiaith ac Acen). Fodd bynnag, dylid nodi bod angen nifer fawr o driffonau, sy'n heriol heb ddata mawr. Er mwyn ymdopi â hyn mewn cyd-destun llai eu hadnoddau fel y Gymraeg (I 2.2, 2.4), gellir clystyru ffonemau sy'n debyg o ran eu nodweddion acwstig. Er enghraifft, gellir clystyru'r seiniau ffrwydrol (*plosive*) /b/ a /p/. Hynny yw, mae nodweddion acwstig y ddau ffonem yn golygu eu bod yn cael eu hynganu mewn ffyrdd tebyg a'u bod yn ystadegol debygol felly o fod yn gyfochrog â llafariad, er enghraifft "beth", "peth". Ceir gwelliant mawr wrth ddefnyddio'r dull clystyru er mwyn amcangyfrif seiniau coll yn iaith Persia yng ngwaith Goodarzi & Almasganj (2016), ac yn ddiweddarach ar gyfer iaith Korea yng ngwaith Bang et al. (2020).

Gwelir isod y broses HMM-GMM yn fras yn Ffigwr 5, lle mae'r GMMs yn dosbarthu nodweddion sydd wedi'u hechdynnu o'r mewnbwn , a'r HMMs yn modelu'r seiniau gyda'u ffonemau cyfatebol. Gelwir y dull hwn yn HMM-GMM yn aml o'r herwydd.



Ffigwr 5 Proses HMM-GMM yn fras

Fel gwelir uchod yn Ffigwr 5, wrth glywed mewnbwn acwstig, mae GMM yn dosbarthu'r nodweddion ac yna, mae'r HMM yn gwirio'r ffonemau mwyaf tebygol. Mae'r model iaith yna'n gwirio'r geiriau mwyaf tebygol. Defnyddir y mewnbwn acwstig 'newyddion' er enghraifft:

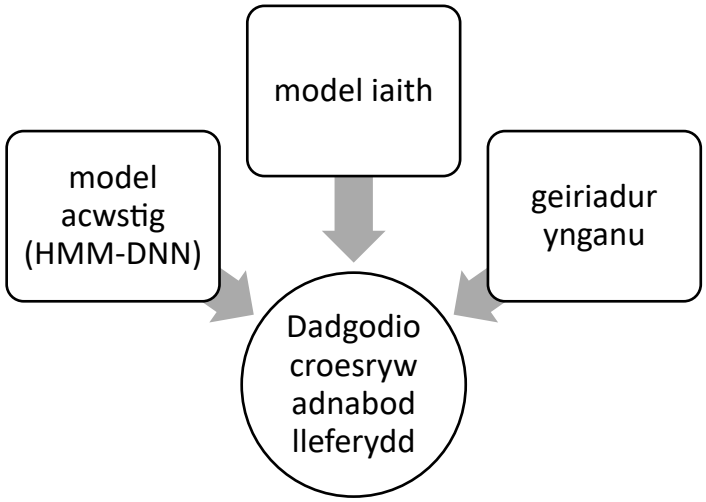
1. dosbarthu nodweddion y mewnbwn acwstig a dadansoddi yn ôl yr isod;
2. ffonem cychwynnol mwyaf tebygol o'r mewnbwn acwstig: 'n';
   ffonem nesaf mwyaf tebygol o'r mewnbwn acwstig: 'EW';
   *HMM yn gwirio bod y dilyniant hwn yn ystadegol debygol;
   ffonem nesaf mwyaf tebygol o'r mewnbwn acwstig: '@' (schwa);
   *HMM yn gwirio bod y dilyniant hwn yn ystadegol debygol;
3. Yna caiff hwn ei wirio ymhellach gan y model iaith, sy'n defnyddio'r geiriadur ynganu i ddadansoddi tebygolrwydd ystadegol dilyniant y geiriau.

Mae'r system yn ailadrodd y broses nes cyrraedd y frawddeg fwyaf ystadegol debygol, er enghraifft "beth yw'r newyddion?".

Gwelir canlyniadau addawol gyda'r dull HMM-GMM gyda'r ieithoedd Hindi yng ngwaith Bansal et al. (2008), yn ddiweddarach ar gyfer Arabeg yng ngwaith Khelifa et al. (2017), ac ar gyfer iaith Indonesia yng ngwaith Hoesen et al. (2016). Fodd bynnag, dengys y gwaith hwn bod mân newidiadau ac addasiadau yn angenrheidiol, neges a gafodd ei nodi'n flaenorol gan Gales a Young (2007). Hynny yw, bod gweithredu systemau gan ddefnyddio'r fath fodelau yn gallu bod yn gymhleth ac yn gofyn am lefel uchel o soffistigedigrwydd. Ar ei orau, wrth i'r fodel HMM-GMM ddysgu dros amser, mae'n medru dadansoddi mewnbwn sain i'r ffonemau mwyaf ystadegol debygol (o'r HMMs), ac yna yn gallu amcangyfrif beth yw'r geiriau mwyaf ystadegol debygol ar sail hynny gyda chymorth y model iaith. Mae modd cyrraedd y nod felly o ddarganfod beth yw'r frawddeg fwyaf ystadegol debygol. Mae'r dulliau ystadegol HMM-GMM yn cael eu defnyddio yng nghit offer HTK yn ogystal ac yng nghydrannau ystadegol cit offer *Kaldi*; ceir mwy am hyn ym mhennod V Citiau Offer.

3.2. Modelau Croesryw

Dros amser, daeth hi'n amlwg nad oedd modd gweld gwelliant mawr wrth ddefnyddio dulliau ystadegol yn unig ar gyfer modelu acwstig. Erbyn y 2010au, amlygir dull newydd o ddadgodio oedd yn gwneud yn fawr o ddatblygiadau newydd ym maes DA: rhwydweithiau niwral (dwfn) ((D)NN). Mae'r dull croesryw yn rhannu'r dulliau ystadegol a rhwydweithiau niwral dwfn, gan gadw'r cydrannau ystadegol HMM ond yn defnyddio cydrannau DNN yn hytrach na rhai GMM er mwyn adnabod monoffon/triffon o nodweddion y sain. Gelwir y dull hwn yn groesiad HMM-DNN felly. Y prif wahaniaeth rhwng model ystadegol fel HMM-GMM a model croesryw fel HMM-DNN yw bod y cyntaf yn dadansoddi bob ffrâm (toriad o sain) ar wahân lle mae HMM-DNN yn fwy hyblyg i allu dadansoddi fframiau gyda'i gilydd (Hinton, et al., 2012). Mae'n cyflymu'r broses ac fel arfer yn gwella'r canlyniadau, fel gwelir yng ngwaith Shahin et al. (2014), Dahl et al. (2012), a Li et al. (2013).



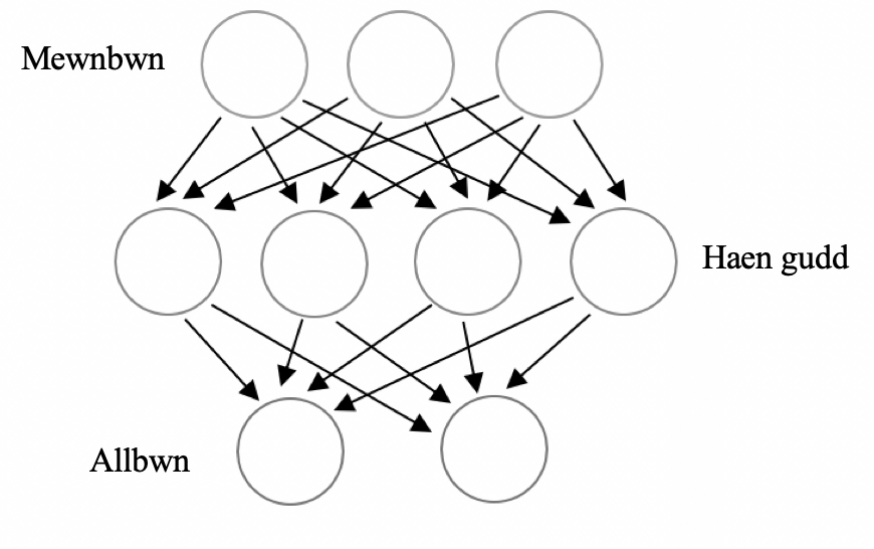
Ffigwr 6 Proses ddadgodio croesryw HMM-DNN adnabod lleferydd yn fras

Mae'r broses ddadgodio fel gwelir yn Ffigwr 5 felly, yn edrych yn debyg ar lefel bras, ond wrth fanylu ar y cydrannau mewnol, gwelir ychwanegiad DNN yn Ffigwr 6. Ceir manylion pellach am bensaernïaeth model DNN isod yn 3.3.

Pŵer model croesryw HMM-DNN yw hyblygrwydd ychwanegol dros HMM-GMM, ond mae gweithredu model o'r fath yn gallu bod yn gymhleth gan bod cyfuniad o gydrannau gwahanol yn gorfod cydweithio â'i gilydd. Fodd bynnag, adlewyrchir poblogrwydd HMM-DNN yng ngwaith Upadhyaya et al. (2017) a Sri et al. (2020) ar gyfer yr ieithoedd Hindi, ac ar gyfer Punjabi yng ngwaith Guglani & Mishra (2018), sy'n gweld canlyniadau addawol. Mae dulliau croesryw HMM-DNN yn cael eu defnyddio ym modelau cit offer *Kaldi* (Povey, et al., 2011); ceir mwy am hyn ym mhennod V.

3.3. Modelau Pen-i-ben

Mae'r modelau diweddaraf yn y maes adnabod lleferydd yn defnyddio modelau pen-i-ben (*end-to-end*) sy'n gwneud i ffwrdd ag HMMs. Maent yn defnyddio rhwydweithiau niwral dwfn ac felly gelwir yn fath o DNN. Mae'r rhain yn gofyn am lai o wybodaeth ieithyddol, sy'n rhoi mwy o hyblygrwydd eto i'r datblygwr. Gwelir pensaernïaeth DNN yn fras isod yn Ffigwr 7. Mae'n dangos sut mae mewnbwn acwstig yn dechrau'r broses, yna mae'r wybodaeth yn cael ei phasio ymlaen i'r haen nesaf, neu haenau nesaf, er mwyn mapio fectorau (*vectors*) o'r sain (fel echdynnu nodweddion uchod, ond gyda rhifau) ac yna ceir yr allbwn fel canlyniad. Gwelir isod (Ffigwr 7) y fersiwn symlaf o'r DNN, lle gall y datblygwr ychwanegu haenau i adlewyrchu lefelau ychwanegol o gymhlethdod yn y model.



Ffigwr 7 Pensaernïaeth rhwydwaith niwral dwfn (DNN) yn fras

(wedi addasu o Meyer (2019, t. 130), lle mae'r rhesi o gylchoedd yn cyfeirio at haenau gwahanol)

Daw'r enw niwral yn wreiddiol o fod wedi seilio ar niwron McCulloch & Pitts (1943), ac mae'n cyfeirio at fodel sydd wedi'i symleiddio o'r niwron mewn ymennydd dynol, ond fel cydran gyfrifiadurol. Er, mae Jurafsky & Martin (2019, t. 123) yn nodi nad yw'r dechnoleg wir yn adlewyrchu hyn rhagor. Yn hytrach, cyflwynir rhwydwaith bwydo ymlaen (*feed forward*), neu *multi-layer perceptrons* (MLP), (Jurafsky & Martin, 2019, tt. 129-132), fel gwelir uchod yn Ffigwr 7, sy'n bwydo gwybodaeth ar bob lefel o un haen i'r haen nesaf, a bod pob uned wedi'i gysylltu gan bâr o haenau cyfochrog. Er bod y cysyniad o basio gwybodaeth drwy rwydwaith aml haenog wedi bodoli ers y 1960au, ni chafodd y term dysgu dwfn ei fathu tan 1980au (Dechter, 1986), sy'n adlewyrchu natur y rhwydweithiau niwral wrth fynd yn ddyfnach (hynny yw, bod mwy ohonynt).

Oherwydd diffyg data (I 2.4) a diffyg pŵer cyfrifiadurol am ddegawdau, doedd dim llawer o frwdfrydedd am y gwaith tan yn gymharol ddiweddar. Ers hynny fodd bynnag, gwelwyd llwyddiant wrth ddefnyddio dulliau DNN ar gyfer adnabod lleferydd, megis gwaith Wang et al. (2019), yn ogystal ag eraill fel gwelir ym mhennod III 2.4 Mozilla, *Common Voice* a *Common Voice-cy*. Dywed Jurafsky & Martin (2019, t. 137) yn glir bod unrhyw faint o ddata yn perfformio'n well o fewn model DNN yn hytrach nac un ystadegol. Ond, maent hefyd yn nodi'r gost am hyn. Hynny yw, eu bod yn arafach i

hyfforddi (2019, t. 138), yn ogystal â'r pŵer cyfrifiadurol a'r data ychwanegol sydd angen. Fodd bynnag, mae peth gwaith wedi dangos llwyddiant wrth geisio cyflymu'r broses (Amodei, et al., 2016). O ganlyniad i ddatblygiadau mewn caledwedd unedau prosesu graffeg (*Graphics Processing Units*, GPUs), mae'r broses wedi cyflymu'n aruthrol, felly'n tynnu sylw cwmnïau mawr y byd megis Google fel gwelir yn nhrosolwg Mittal & Vaishnay (2019). Erbyn hyn, mae ymchwil gan dimau ymchwil megis Agarwal a Zesch (2020) (71.5% WER ar <100 awr Almaeneg y Swistir (isaf y WER y gorau oll yw'r canlyniad)), a Hjortnæs et al. (2020) (99.1% WER ar 35 awr Komi), yn cydnabod cyfyngiadau DNN, a model pen-i-ben yn benodol, ar gyfer ieithoedd llai eu hadnoddau, felly'n edrych i ddulliau eraill am gymorth (mwy yn 4). Mae'r model pen-i-ben yn cael ei ddefnyddio yn *DeepSpeech* (Ardila, et al., 2020) er enghraifft, ac fe geir mwy am hyn ym mhennod V Citiau Offer. Fel gwelir ym mhennod VIII *DeepSpeech*, mae peth ymchwil ar ddefnydd model pen-i-ben, a DNN yn gyffredinol, o fewn ieithoedd llai eu hadnoddau fesul iaith unigol ond ddim ar gyfer y Gymraeg ar adeg ysgrifennu.

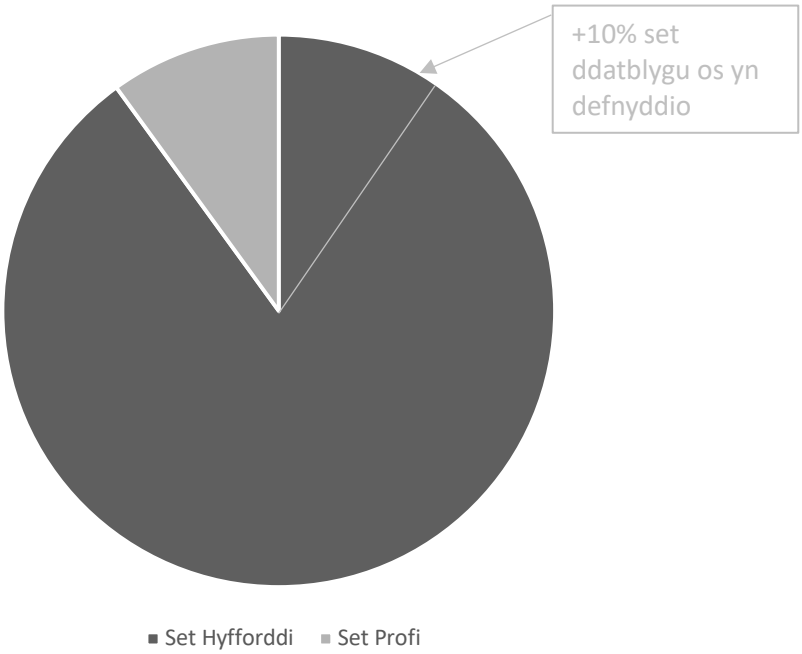
4. Dulliau i Ymdopi gyda Llai o Ddata

Gan fod diffyg adnoddau yn her wrth ddefnyddio'r dulliau mwyaf diweddar a welir yn 3.3, mae ieithoedd llai eu hadnoddau dan anfantais. Mae timau ymchwil megis Baetev et al. (2018), Toshinwal et al. (2018) a Cui et al. (2015), yn troi at ddulliau eraill (isod) i ymdopi â hyn felly. Pwrpas y rhan hon yw rhoi cyflwyniad i rai dulliau eraill sydd wedi profi'n addawol yng nghyd-destun diffyg data, her gyffredin ymysg ieithoedd llai eu hadnoddau fel nodir uchod, ac yn 1.1. Yn gyntaf, ceir cyflwyniad i'r dull k-*fold cross validation* (C-V), lle mae modd rhannu'r data fewn i nifer o rannau bach er mwyn eu hail-ddefnyddio a gwneud yn fawr o'r corpora cyfan. Yn ail, ceir cyflwyniad i drosglwyddo dysgu, fel nodir uchod, wrth gofio bod trafodaeth ehangach o oblygiadau hyn ym mhennod VIII 4 Trafodaeth.

k-*fold cross-validation*

Ceir enghraifft o'r dull hwn ar waith ym mhennod V 2.2 Gweithdrefn. Mae'n ddull sy'n galluogi defnydd o gorpws llawn fel nad oes angen neilltuo data ar gyfer profi yn unig (esboniad isod). Yn ôl Jurafsky & Martin (2019, t. 36), y ffordd orau o asesu perfformiad system adnabod lleferydd yw'r asesiad anghynhenid (*extrinsic evaluation*). Ond, maent yn pwysleisio bod y dull hwn yn llafurus ac araf, ac felly'n awgrymu defnyddio'r asesiad

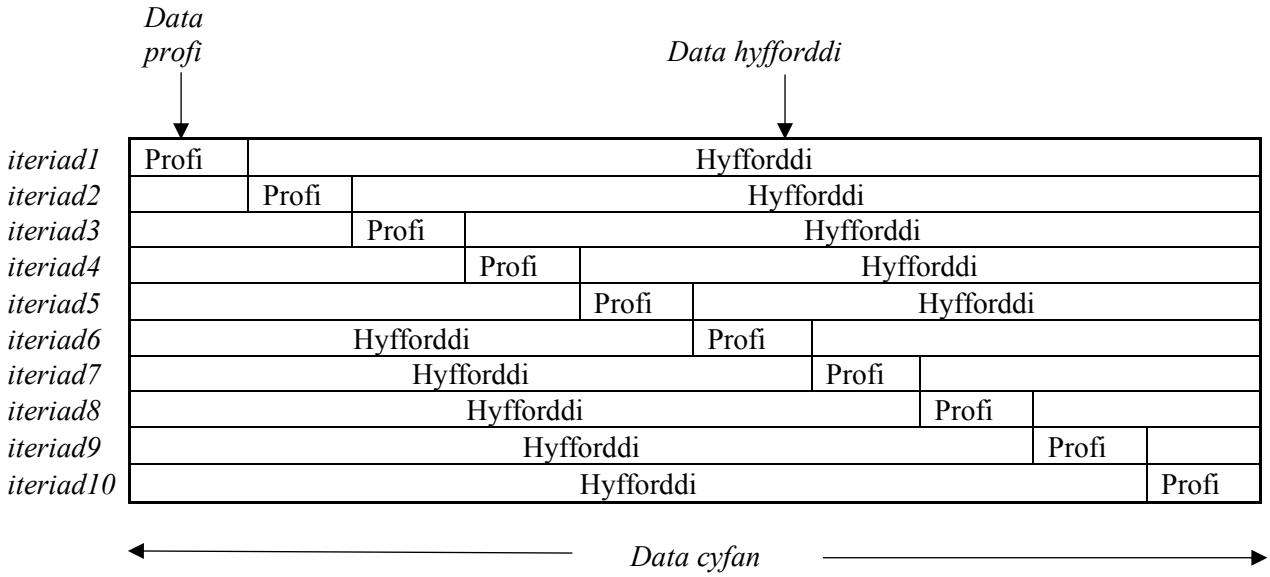
cynhenid (*intrinsic evaluation*) yn ei le (2019, t. 36). Gwelir isod ddyluniad o sut defnyddir y data er mwyn gweithredu'r asesiad cynhenid (Ffigwr 8).



Ffigwr 8 Asesiad cynhenid

Yn arferol, rhennir y data i set hyfforddi (*training set*) (90%) a set brofi (*test set*) (10%) (a set ddatblygu ambell waith, esboniad yn 2.2 Gweithdrefn). Y gwendid yma ar gyfer ieithoedd llai eu hadnoddau yw bod angen neilltuo 10% o'r data ar gyfer y set brofi. Wrth ymdopi â diffyg data, gall neilltuo 10% fod yn aberth enfawr (Jurafsky & Martin, 2019, t. 69) sy'n gallu arwain at arferion gwael (mwy ym mhennod V 2.2 Gweithdrefn). Dylid nodi bod ystyriaeth ychwanegol yma sef effaith ymdebygu setiau hyfforddi a phrofi. Mae Kleynhans & Barnard (2015) yn nodi'n glir y bydd system adnabod lleferydd yn perfformio'n wael os yw'r set hyfforddi a'r set brofi yn hollol wahanol. Caiff hyn ei atgyfnerthu gan waith Goyvêa & Davel (2011) sy'n dangos bod cynllunio data tebyg yn gwella perfformiad. Mae gwaith Koenecke et al. (2020) yn adleisio hyn yn eu gwaith ar wahaniaethau rhwng hil wrth gyfathrebu gyda dyfeisiau adnabod lleferydd. Wrth ddatblygu corpws o siaradwyr gwyn a du o bum dinas Americanaidd, oedd wedi'i reoli am amrywiaethau oedran a chenedl er enghraifft, cyfrifir WER o 0.35% ar gyfer siaradwyr du, a 0.19% ar gyfer siaradwyr gwyn. Eu hawgrym i wella hyn oedd adlewyrchu'r amrywiadau hyn yn y data hyfforddi, wrth greu corpora mwy amrywiol.

Un datrysiad i hyn oll yw defnyddio'r dull k-*fold cross-validation* (C-V). Mae'n ddull effeithiol a dibynadwy sy'n cael ei ddefnyddio'n aml gan y gymuned dysgu peirianyddol (Jurafsky & Martin, 2019, t. 69), (Cawley & Talbot, 2010, t. 2082). Gwelir enghraifft ohono ar waith yn arbrofion Caon et al. (2011), sy'n dangos gwelliant wrth ddefnyddio C-V ar gyfer modelau Ffrangeg. Mae'r dull yn defnyddio paramedr sengl o'r enw k sy'n cyfeirio at faint o blygiadau (*folds)* fydd yn cael eu creu o'r set data cyfan (hynny yw, faint o raniadau). Os yw k=10, rhennir y data i 10 plyg ar hap, ac yna bob plyg i set hyfforddi (90%) a set brofi (10%). Wrth redeg 10 prawf felly, yn defnyddio plyg gwahanol bob tro, mae modd cyfrifo 10 canlyniad a chymryd cyfartaledd, heb orfod neilltuo unrhyw ddata. Gwelir sut defnyddir y data isod yn Ffigwr 9.



Ffigwr 9 k-*fold cross-validation*

Trosglwyddo Dysgu

Pwrpas y rhan hon yw rhoi cyflwyniad i'r dull trosglwyddo dysgu. Mae'n ddull lled-ddiweddar ymysg ieithoedd llai eu hadnoddau sy'n cynnig opsiynau newydd i'r gymuned, felly caiff ei drafod ymhellach yng nghyd-destun gwaith pellach posib ym mhennod IX 6 Ymchwil Pellach. Mae'n ddull lle defnyddir data o iaith (neu ieithoedd) fwy ei hadnoddau er mwyn hyfforddi system adnabod lleferydd cyn ychwanegu iaith lai ei hadnoddau ar ddiwedd y broses, yn manteisio felly ar ddata yr iaith fwy. Gan ddefnyddio'r dull trosglwyddo dysgu, cyfrifwyd canlyniadau addawol o fewn systemau pen-i-ben (3.3) ar gyfer Twrcaidd (80 awr, 45.8% WER, (Baetev et al., 2018)), ar gyfer

ieithoedd India (1590 awr, 22.91-29.05% (Toshinwal, et al., 2018)) a Swahili (3 awr, tua 60% WER (Cui, et al., 2015)). Dylid nodi bod rhai cyfyngiadau i effeithiolrwydd y dull, ac fe gaiff y rhain eu trafod ymhellach ym mhennod VIII 4 Trafodaeth. Dylid nodi, er enghraifft, bod ieithoedd sydd â rhai tebygolrwyddau ieithyddol (mwy ym mhennod IV Cefndir: Amrywio yn y Gymraeg) yn dueddol o weithio'n well gyda'i gilydd. Gwelir enghraifft o hyn gan Kunze et al. (2017), lle defnyddir data bach Almaeneg ar system wedi'i hyfforddi ar ddata Saesneg. Enghraifft o waith tebyg arall yw Barroso et al., (2012) lle dengys bod modd tynnu ar adnoddau Ffrangeg a Sbaeneg er mwyn datblygu technoleg lleferydd Basgeg. Daw'r syniad yn wreiddiol o weithiau megis Yarowsky et al. (2001) lle defnyddir adnoddau Saesneg i hyfforddi corpws dwyieithog. Enghreifftiau eraill yw gwaith a wnaed yn y 1990au gan Köhler (1996), a gan Schultz & Waibel (2001) er enghraifft, a ddefnyddiodd dulliau trosglwyddo dysgu i adeiladu modelau acwstig amlieithog. Maent ill dau yn gweld ychydig o welliant yn eu canlyniadau wrth ddefnyddio ieithoedd tebyg i'w gilydd fel iaith hyfforddi a tharged, er enghraifft Saesneg ac Almaeneg, yn y ddau achos. Fodd bynnag, mae'n ddull sy'n dal i gael ei ddatblygu fel noda Chen et al. (2019). Mae gwaith diweddar Meyer (2019) hefyd yn cyfrifo canlyniadau addawol ar gyfer nifer o ieithoedd llai eu hadnoddau gan ddefnyddio corpws *Common Voice* o fewn peiriant *DeepSpeech*, ond eto yn tynnu sylw at rai cyfyngiadau fel nodir uchod. Trafodir goblygiadau hyn ymhellach ym mhennod VIII 4 Trafodaeth, ac fel datblygiad posib ar gyfer y dyfodol ym mhennod IX 6 Ymchwil Pellach.

5. Casgliadau

Nod y bennod hon oedd cynnig cyflwyniad i gefndir y maes technoleg lleferydd, cyn edrych yn fanylach ar ddefnydd ymarferol y modelau gwahanol ar gyfer adnabod lleferydd Cymraeg yn benodol.

Gwelwyd yn 2 bod LVCSR (adnabod lleferydd geirfa eang di-dor) yn ffocws yn y traethawd hir hwn. Yn ogystal â rhannu technoleg sydd o gymorth i'r meysydd archifo, addysg, ac iechyd, fel gwelwyd yn 1.2, mae'n ddull cyffroes er mwyn cyfathrebu gyda dyfeisiau personol wrth ddefnyddio cynorthwywyr deallus er enghraifft (I 2 Cyd-destun Ymchwil). Fel nodir ym mhennod I Cyflwyniad, mae'r Gymraeg yn iaith lai ei hadnoddau sy'n dioddef o ddiffyg data ac felly rhaid ystyried y gwahanol gyfyngiadau wrth addasu er mwyn defnyddio technegau mwy diweddar.

Cafwyd cyflwyniad i dri math o ddull yn 3: model ystadegol (gan ddefnyddio modelau ystadegol ar gyfer modelu acwstig), model croesryw (dull ystadegol ond gan

ychwanegu dulliau niwral ar gyfer modelu acwstig), a model pen-i-ben (defnyddio dulliau niwral yn unig ar gyfer cyflawni'r dasg gyfan). Mae'r ddau gyntaf yn defnyddio elfennau o fodelau ystadegol i wahanol raddau, lle mae datblygiad pen-i-ben yn defnyddio dulliau mwy diweddar rhwydweithiau niwral dwfn. Cafwyd esboniad o hyn yn 3.3, lle gwelir dyluniad o bensaernïaeth y system wedi'i symleiddio (Ffigwr 7). Gellir cymharu'r dyluniad hwn gyda'r ffigwr cyfatebol ystadegol yn Ffigwr *5*. Roedd 4 yn gyflwyniad i rai dulliau defnyddiol i ymdopi â diffyg data ymysg ieithoedd llai eu hadnoddau. Mae manylion pellach am ym mhenodau V 2.2 Gweithdrefn a VIII 4 Trafodaeth, ond i grynhoi, tra bod C-V (k-*fold cross-validation*) yn ddull ymarferol ac effeithiol o wneud yn fawr o ddata llai, mae trosglwyddo dysgu i'w weld yn faes cyffroes yn nyfodol y maes niwral.

Mae ymchwil wedi dechrau ymddangos sy'n defnyddio modelau niwral mewn amgylcheddau heriol (Watanabe et al., 2017). Fel iaith lai ei hadnoddau, mae'r Gymraeg mewn sefyllfa heriol yn barod, felly mae angen ystyried cyfyngiadau megis maint data a phŵer cyfrifiadurol (Jurafsky & Martin, 2019, t. 138), cyn dechrau addasu dulliau i rai mwy diweddar. Rhai o brif amcanion y traethawd hir hwn yw ymchwilio i fathau gwahanol o fodelau a meintiau gwahanol o ddata Cymraeg er mwyn adeiladu system gadarn. Fel nodwyd yn 1, gwelir y cwestiynau ymchwil sy'n arwain ymlaen o hyn oll ym mhennod IV 5 Amcanion yr Ymchwil. Cyn hynny, yn y bennod nesaf, ceir trosolwg o'r gwaith sydd wedi'i gyflawni o fewn y maes hwn ar gyfer y Gymraeg cyn belled; gwaith sydd wedi gwneud yn fawr o'r dulliau ystadegol a chroesiad a gyflwynir yn 3, ac ystyriaeth o enghreifftiau gwahanol o amrywio yn y Gymraeg ym mhennod IV Cefndir: Amrywio yn y Gymraeg.

III. Cefndir: Technoleg Lleferydd Cymraeg

1. Cyflwyniad

Bydd y bennod hon yn rhoi trosolwg o'r gwaith sydd wedi'i gyflawni, a'r adnoddau sydd wedi'u creu yn barod ar gyfer technoleg lleferydd Cymraeg. Ceir cyflwyniad i'r cyd-destun llai ei hadnoddau yn 1.1, fel a ddiffinnir ym mhennod I 2.2 'Iaith Lai ei Hadnoddau', a'r heriau cysylltiedig, cyn manylu ar her ychwanegol trwyddedu yn 1.2. Ceir ystyriaeth fanylach o ddiffyg data oherwydd amrywio yn yr iaith ym mhennod IV Amrywio yn y Gymraeg. Yn y bennod hon, mae 2 yn rhoi trosolwg o ddatblygiad technoleg lleferydd y Gymraeg a'r projectau mwy diweddar sy'n arwain at adnabod lleferydd ar gyfer cynorthwywyr deallus yn 2.4.

Mae'n arwain at y bennod nesaf, sy'n edrych yn fwy penodol ar agweddau heriol i brosesu yn yr iaith Gymraeg, IV Amrywio yn y Gymraeg. Gwelir yr amcanion a chwestiynau ymchwil sy'n arwain ymlaen o hyn oll ar ddiwedd pennod IV 5 Amcanion yr Ymchwil.

1.1. Ieithoedd Llai eu Hadnoddau a Diffyg Data

Fel nodir uchod, ceir cyflwyniad bras i ieithoedd llai eu hadnoddau ym mhennod I 2.2 'Iaith Lai ei Hadnoddau'. Yna, diffinnir y label ei hun, ac fe esbonnir nad oes digon o ddata i ddatblygu'r dechnoleg, dim llawer o gymorth ariannol i gasglu data, neu i gyflogi datblygwyr neu arbenigwyr ieithyddol, ac yn y blaen. Dylid nodi eto, nad yw pob iaith lai ei hadnoddau yn lleiafrifol. Er enghraifft, mae Sindhi yn iaith lai ei hadnoddau (Jamro, 2017), er ei bod hi'n cael ei siarad gan dros 75 miliwn o bobl dros Bakistan ac India, ac â dros 40 miliwn o siaradwyr ym Mhakistan yn unig. Mae ieithoedd llai eu hadnoddau fel Sindhi yn tueddu i fod yn ieithoedd cymunedau tlawd a difreintiedig, neu'n rhai nad ydynt eto wedi mabwysiadu technolegau modern yn llawn. Maent yn dioddef o ddiffyg data a diffyg arbenigwyr yn gweithio arni. Mae peth ymchwil wedi'i gwblhau a rhai adnoddau wedi'u creu ar gyfer Sindhi, fel gwelir yn nhrosolwg Jamro (2017) ond mae'n nodi nad yw'r rhain ar gael yn gyhoeddus sy'n ychwanegu at yr her (mwy isod yn 1.2). Enghraifft arall yw Arabeg, sy'n un o ieithoedd mwyaf y byd gyda thros 300 miliwn o bobl yn ei siarad fel mamiaith (Khelifa et al., 2017, t. 942). Er hyn, mae'n dioddef o ddiffyg data addas ar gyfer hyfforddi a phrofi modelau ar gyfer technoleg lleferydd fel gwelir yng ngwaith Khelifa et al. (2017). Mae gwaith Khelifa et al. (2017) yn dogfennu eu hymdrechion wrth ddatblygu corpws 8 awr o hyd gan 10 siaradwr er mwyn hyfforddi

modelau ar gyfer adnabod lleferydd Arabeg. Er mwyn deall y cyd-destun, gwelir yng ngwaith Prabhavalkar et al. (2017, t. 939) bod ymchwil gan Google a NVIDIA, cwmnïoedd technolegol rhyngwladol, yn defnyddio corpws mawr iawn, o tua 12,500 awr, er mwyn gwneud cymhariaeth o fodelau addas ar gyfer adnabod lleferydd. Maent yn darganfod bod maint mor fawr o ddata yn arwain at ganlyniadau da iawn sy'n cystadlu gyda'r systemau gorau ar y pryd. Mae corpora eraill ar gael yn gyhoeddus ar gyfer Saesneg, er enghraifft TED-LIUM (Hernandez et al., 2019) a *Librispeech* (Panayotov et al., 2015), sydd rhyngddynt yn rhoi cyfanswm o tua 1500 awr o ddata sain. Dylid pwysleisio nad yw'r rhain yn cynnwys corpora preifat cwmnïoedd masnachol Apple, Amazon neu Google er enghraifft.

Mae'r Gymraeg yn iaith leiafrifol sydd hefyd yn llai ei hadnoddau. Yn ôl Arolwg Blynyddol o'r Boblogaeth rhwng Ebrill 2019 a Mawrth 2020 (2020), roedd 28.3% o bobl tair oed a throsodd yng Nghymru yn gallu siarad Cymraeg. Mae'r ffigur hwn yn cyfateb i 855,200 o bobl. Ceir isod yn 2, drosolwg manylach o'r adnoddau sydd wedi'u datblygu ar gyfer yr iaith, gan gynnwys system gwirio gramadeg, lemateiddiwr, a gwaith ar gyfer adnabod lleferydd, ar gyfer cynorthwyydd personol deallus yn fwy diweddar. Ar adeg profi y traethawd hir hwn, roedd 80 awr o ddata sain addas ar gael yn y Gymraeg (wedi ei seilio ar ffigurau corpora Paldaruo 4.0 (Cooper et al., 2018) a *Common Voice-cy* (13/02/19, fersiwn cynnar cyn cyhoeddi swyddogol) (Mozilla, 2020)), ac mae Tabl 1 yn crynhoi'r ffigurau uchod i dabl. Gwneir hyn er mwyn adlewyrchu'r gwahaniaeth mawr sydd i'w gael rhwng data addas mewn ieithoedd gwahanol. Ceir enghraifft ychwanegol o iaith leiafrifol sy'n llai ei hadnoddau isod hefyd, sef Ffrisieg gyda chorpws FAME (Yılmaz, et al., 2016).

	Corpws	Oriau
Cymraeg	Paldaruo 4.0	38
	Common Voice-cy (13/02/19)	42
Saesneg	TED-LIUM	452
	Librispeech	tua 1,000
Enghreifftiau eraill	Google/NVIDIA (uchod)	tua 12,500
	Arabeg (uchod)	8
	Ffrisieg (FAME)	14

Tabl 1 Cymharu data sain addas ar gyfer adnabod lleferydd Cymraeg a Saesneg, ac eraill

Fel nodir uchod, mae'r Uned Technolegau Iaith, Canolfan Bedwyr, wrth gydweithio gydag ieithyddion ym Mhrifysgol Bangor, yn cyhoeddi gwaith ymchwil ac adnoddau arbenigol ar gyfer technoleg lleferydd megis corpws Paldaruo (Cooper et al., 2018) a rheolau llythyren-i-sain (*letter-to-sound rules*, LTS (Chan & Cooper, 2014)). Mae'r gwaith yma yn cael ei gyhoeddi yn god agored sy'n allweddol ar gyfer ieithoedd llai eu hadnoddau (mwy isod yn 1.2). Yn ddelfrydol, byddai modd defnyddio corpws o ddata sy'n barod wedi'i gasglu (fel, i raddau, yn Jones et al. (1998) lle casglwyd data o gorpora BT a straeon newyddion Cymraeg), ond gall ffactorau fel, er enghraifft, costau trwydded fod yn rhwystr i gasglu data ac yn y blaen.

1.2. Trwyddedu

Ceir uchod rhai enghreifftiau o ieithoedd sy'n dioddef o ddiffyg adnoddau data, lle sonnir yn fras am her ychwanegol trwyddedu. Diffyg cyllid yw'r her sydd ynghlwm â thrwyddedu gan fod angen prynu mynediad at rai corpora caeedig, lle gall costau amrywio'n fawr. Fel awgrymir uchod, mae cewri masnachol y maes megis Apple, Amazon, a Google er enghraifft, yn manteisio ar allu talu am, neu ddatblygu adnoddau mawr all fod dan drwyddedi caeedig. Heb gefnogaeth ddigonol ar gyfer ieithoedd llai eu hadnoddau, mae'n her ychwanegol sy'n wynebu'r ieithoedd hynny sydd heb yr adnoddau addas.

Mae trwyddedu yn rheoli rhyddid datblygwyr i ddefnyddio data at eu dibenion eu hunain. Oherwydd y cyfyngiadau sydd ar ddata caeedig (nid yn agored hynny yw), mae'n

anodd ei ddefnyddio er mwyn hyfforddi system adnabod lleferydd er enghraifft. Mae sawl math o drwyddedau ar gael; er enghraifft, perchnogol (*proprietary*, sy'n gaeedig) neu ganiataol (*permissive*, sy'n agored). Gyda thrwyddedi agored, does dim cyfyngiadau neu ymrwymiadau rhwystredig i'r datblygwr. O dan drwydded agored, mae modd defnyddio'r cynnyrch at unrhyw bwrpas: ei ddosbarthu, ei olygu ac yna i ddosbarthu'r fersiynau wedi diwygio heb ystyried breindaliadau. Enghraifft o hyn yw beth mae Mozilla yn defnyddio yn eu project *Common Voice* (isod) sef *Creative Commons* (CC0). Eto, mae'n agored ac nid oes cyfyngiadau ar y defnydd ohono (Choukri, et al., 2012, tt. 40-45). Mae trwyddedau caeedig ar y llaw arall, fel sydd gan y *Linguistic Data Consortium* ymhlith eraill (Choukri, et al., 2012, tt. 40-45), yn dynn iawn am ailddosbarthu eu cynnyrch. O fewn trwyddedau perchnogol er enghraifft, mae modd i'r perchennog gyfyngu defnydd, golygu ac ailddosbarthiad y cynnyrch. Rhai enghreifftiau yw trwyddedau ar gyfer cynnyrch *Microsoft Word*, *Adobe Flash Player*, *iTunes*, *macOS* a *Skype.*

Mae'r is-ran hon wedi'i gynnwys yn y bennod hon er mwyn rhoi syniad o'r sefyllfa drwyddedu, all fod yn her ychwanegol mewn cyd-destun iaith lai ei hadnoddau. Lle nad oes modd manteisio ar gorpora sy'n bodoli yn barod, unai am resymau argaeledd data neu ddiffyg adnoddau cyllid i brynu mynediad, gellid troi at y broses heriol o gasglu data newydd. Ceir isod yn 2, drosolwg o'r gwaith hwn ar gyfer y Gymraeg.

2. Technoleg lleferydd Cymraeg

Pwrpas yr is-bennod hon yw rhoi trosolwg o'r gwaith sydd wedi'i gyflawni a'r adnoddau sydd wedi'u creu ar gyfer neu sy'n cyfrannu at dechnoleg lleferydd Cymraeg, gan gynnwys system gwirio gramadeg CySill, LTS (rheolau llythyren-i-sain) a chorpora gwahanol, projectau ar y cyd gydag ieithoedd eraill, a gwaith tuag at gynorthwyydd personol deallus Cymraeg. Mae'r *Digital Language Survival Kit* (Berger, et al., 2018, t. 10) yn nodi bod datblygu'r adnoddau hyn yn hanfodol ar gyfer adferiad iaith. Maent yn cynnwys geiriaduron electronig, system gwirio gramadeg, corpora digidol o feintiau gwahanol, ac adnabod lleferydd yn eu rhestr o adnoddau hanfodol. Wrth drafod adnodd adnabod lleferydd, maent hefyd yn cydnabod bod y rhan fwyaf o dimau ymchwil yn dechrau wrth ddatblygu synthesis lleferydd cyn adnabod lleferydd, tasg debyg sy'n rhannu rhai o'r un cydrannau ond sy'n ddynwarediad cyfrifiadurol o lais dynol, yn hytrach na'i hadnabod a'i phrosesu (Berger, et al., 2018, t. 21). Caiff hyn ei adlewyrchu

yn y gwaith sydd wedi'i gyflawni a'r adnoddau sydd wedi'u creu ar gyfer y Gymraeg isod.

O ddeall yr heriau sy'n wynebu ieithoedd llai eu hadnodau o 1 uchod, trodd tîm ymchwil yr Uned Technolegau Iaith, Canolfan Bedwyr, at y broses heriol o ddatblygu rhai o'r adnoddau hyn, fel gwelir isod. Dyma brif ddatblygwr deunyddiau electronig ar gyfer yr iaith Gymraeg erbyn heddiw (Prys, 2006, t. 50), (Prys, 2008, t. 35).

2.1. Geiriadur, CySill a Dechrau Casglu Data

Ym Mangor, datblygwyd un o ddarnau cyntaf offer technoleg lleferydd Cymraeg yn adrannau Seicoleg y Brifysgol: CySill (CYmorth SILLafu) (Ellis et al., 2000), sef system gwirio gramadeg Cymraeg ar gyfer *Microsoft Windows*. Cafodd hwn ei ddiweddaru a'i hail-ysgrifennu fel Cysill (yn hytrach na CySill felly) gan Brifysgol Bangor yn 2004 a'i becynnu gyda geiriaduron electronig fel rhan o Cysgliad (Prys et al., 2016). Gwelir isod yn 2.2 sut arweiniodd hyn at gasgliad y corpws testun mwyaf ar gyfer y Gymraeg, sy'n adnodd hollbwysig ar gyfer adferiad yr iaith (Berger, et al., 2018).

Ym Mhrifysgol Caeredin, yn y Ganolfan ar gyfer Ymchwil Technoleg Lleferydd (*The Centre for Speech Technology Research, CSTR*), roedd ymchwil yn cael ei gynnal ar greu syntheseiddydd testun-i-leferydd (TIL) Cymraeg (Williams, 1994), (Williams, 1995). Roedd y system gychwynnol yn cynnwys 51 ffonem, gan gynnwys 3 sain Saesneg er mwyn gallu adnabod enwau llefydd Saesneg sy'n ymddangos yn aml mewn lleferydd Cymraeg. I ddechrau, crëwyd rhestr o eiriau a brawddegau diystyr Cymraeg, yn defnyddio'r orgraff arferol, gan fod sillafu Cymraeg mor dryloyw ac felly'n adlewyrchu'r ynganiad gan amlaf. Recordiwyd siaradwr Cymraeg gwrywaidd o'r de yn darllen o'r rhestr, yna crëwyd mynegai o'r geiriau a geiriadur cyfatebol drwy gyfuniad o segmentu â llaw ac awtomatig.

Rheolau llythyren-i-sain

Er mwyn cwblhau'r system TIL, ysgrifennwyd fersiwn o'r LTS Cymraeg (Williams, 1994). Pwrpas LTS Williams (1994) oedd ceisio galluogi syntheseiddydd TIL i adnabod geiriau o du allan i'r geiriadur ynganu (Williams, 1994, t. 741). Fel nodir yn y bennod flaenorol, mae'r geiriadur ynganu yn gydran ieithyddol bwysig o'r system, lle ceir cofnod o'r holl eiriau mae'r system yn gallu adnabod (Williams, 1994, t. 261). Mae'r broses yma yn trosi fersiwn orgraffyddol gair i'w drawsgrifiad ffonemig (Williams, 1994, t. 261). Yn LTS Williams (1994), roedd modd cydnabod llafariaid epenthetig hefyd, er enghraifft,

‘cefn’ /kévɜn/, ac enghreifftiau eraill megis ‘heol’ /héɔl/, yn ogystal â phwyslais, sy’n cyfoethogi’r ystod o eiriau gwahanol all gael eu prosesu’n gyfrifiadurol.

Mae’n ddiddorol nodi bod Williams wedi awgrymu, yng nghanol y 1990au (1994), (1995), mai’r cam nesaf fyddai ehangu’r syntheseiddiwr i allu prosesu acen y gogledd, agwedd sy’n cael ei ymchwilio yn y traethawd hir hwn ym mhennod VII Amrywio Iaith. Ym 1999, datblygwyd corpws bach ar gyfer y Gymraeg lle’r oedd is-set lai wedi’i anodi â llaw (Williams, 1999) fyddai’n cyfrannu at y gwaith o greu llais synthetig gwell ar gyfer y Gymraeg.

SpeechDat Cymru

Ym 1998-99, cyhoeddwyd SpeechDat Cymru (Jones et al., 1998), oedd yn gorpws o 2,000 siaradwr ar gyfer adnabod lleferydd Cymraeg. Cafodd ei recordio dros rwydwaith ffôn fel rhan o SpeechDat(II), project ELRA (*European Languages Resources Association*) oedd wrthi’n casglu data lleferydd mewn dros 20 o ieithoedd Ewropeaidd gwahanol. Pwrpas y project Cymraeg oedd cynnig cyfleoedd newydd i ddatblygu technoleg ar gyfer yr iaith, ar gyfer adnabod lleferydd ond hefyd ar gyfer ymchwil ym meysydd seineg, synthesis lleferydd ac ieithyddiaeth yn gyffredinol. Yn fwy na dim, mae gwaith Jones et al. (1998), (2000) yn pwysleisio sut mae rôl y Gymraeg mewn projectau rhyngwladol o’r fath yn codi statws yr iaith i lwyfan sylweddol fwy. Wrth gasglu’r data, gwnaed ymdrech arbennig i ehangu’r cyfyngiadau ar argaeledd data addas, fel noda Williams (1999). Yn benodol, gwnaed ymdrech i gasglu geiriau a brawddegau morffolegol gyfoethog er mwyn sicrhau adlewyrchiad o’r iaith gyfan. Daw hwn yn fwy perthnasol fyth gyda datblygiadau pellach yn y maes adnabod lleferydd, megis trosglwyddo dysgu. Gwnaed ymdrech hefyd i sicrhau dosbarthiad cytbwys o genedl enwau (45/55% gwryw/benyw) ac oedrannau (gweler Jones et al. (1998)). Noda Jones et al. (1998, tt. 2-6) y prif heriau ieithyddol a wynebai’r iaith bryd hynny hefyd, sy’n dal i wynebu’r iaith heddiw fel nodir ym mhennod IV Amrywio yn y Gymraeg. Wrth drafod acen, mae sôn am amrywio mewn hyd llafariaid ac ynganiadau gwahanol eraill, amrywio wrth ddefnyddio’r system dreiglo ac amrywiadau mewn arddull hefyd. Llwyddodd SpeechDat Cymru i godi ymwybyddiaeth o’r iaith ymysg y gymuned Ewropeaidd.

2.2. WISPR, Cysill Ar-lein a Lemateiddiwr

Er mwyn mynd a’r gwaith uchod ymhellach, sefydlwyd y project WISPR (*Welsh and Irish Speech Processing Resources*, (WISPR, diddyddiad)). Roedd WISPR yn broject

gafodd ei ariannu gan y rhaglen *Interreg* Ewrop a Bwrdd yr Iaith Gymraeg. Bwriad y project oedd datblygu synthesis TIL ar gyfer y Gymraeg a Gwyddeleg, wrth gasglu data sain ar gyfer yr ieithoedd hynny (Williams et al., 2006). Cafodd ei ddatblygu ar y cyd gan yr Uned Technolegau Iaith, Canolfan Bedwyr, Prifysgol Bangor, a Choleg y Drindod yn Nulyn, gyda chefnogaeth gan Brifysgol Dinas Dulyn a Choleg Prifysgol Dulyn. Roedd aelodau'r tîm WISPR yn pwysleisio mai'r ffordd gorau i ledaenu adnabod lleferydd iaith lai ei hadnoddau oedd darparu offer oedd yn caniatáu ailddosbarthiad rhydd (1.2). Roeddent yn pwysleisio y dylid cael trwydded syml a hael, er mwyn caniatáu datblygwyr i'w ddefnyddio a'i integreiddio fewn i'w meddalwedd eu hunain. Mae'r holl ffynonellau, meddalwedd ac offer a ddatblygwyd yn ystod y project WISPR ar gael am ddim ar gyfer unrhyw bwrpas. Mae hwn yn sicrhau cyn lleied o gyfyngiadau a phosib, fel bod allbwn technoleg lleferydd WISPR yn medru cael ei ddefnyddio gydag ystod eang o feddalwedd, yn cynnwys meddalwedd perchnogol a chod agored. Rhannwyd y dulliau rhwng y ddwy iaith er mwyn helpu ei gilydd ond wrth gydnabod bod gan Wyddeleg system ffonolegol ac orgraffyddol mwy cymhleth na'r Gymraeg (Prys, et al., 2004, t. 68). Byddai gwaith WISPR ar syntheseiddiwr TIL yn arwain at waith pellach ar dechnoleg adnabod lleferydd Cymraeg.

Yn y cyfamser, cyhoeddwyd Cysill Ar-lein fel ffordd o gynnig y gwasanaeth gwirio, ac adeiladu corpws digidol o Gymraeg gyfoes ysgrifenedig ar yr un pryd (Prys et al., 2016). Roedd y fersiwn ar-lein yn gofyn i ddefnyddwyr gytuno i'r data gael ei ychwanegu at y corpws hwn, wrth gynnig gwasanaeth iddynt oedd yn gwirio 3,000 llythyren ar y tro. Gwelir poblogrwydd y rhaglen wrth i'r fersiynau gael eu defnyddio mewn ysgolion a cholegau, gan unigolion, gan Cynulliad Cenedlaethol Cymru (bellach Senedd Cymru) ac awdurdodau lleol (Hicks, 2004, t. 2). Dengys ymchwil Wooldridge (2011) cymaint oedd ysgolion yn eu defnyddio wrth ddangos gostyngiad yn y defnydd yn ystod gwyliau ysgol. Erbyn 2015, roedd 31 miliwn gair yn y corpws Cysill ar-lein, gan 600 defnyddiwr bob dydd, a thua 1,100 ymweliad i'w wefan yn ddyddiol (Prys et al., 2016, t. 3263). Wrth i fwy o bobl ddefnyddio'r system, roedd y corpws testun yn casglu mwy o enghreifftiau, oedd yn ychwanegu at faint yr adnodd. Yn 2019, roedd y corpws yn cynnwys 150 miliwn o eiriau, sef y corpws mwyaf ar gyfer y Gymraeg ar y pryd (Jones et al., 2019, tt. 18-19). Mae'r fersiwn ar-lein (isod) felly yn cynrychioli un o'r corpora digidol Cymraeg mwyaf, wedi'i ddatblygu'n gost isel, sy'n adnodd arbennig o werthfawr ar gyfer iaith lai ei hadnoddau. Mae'n gynnyrch sy'n adlewyrchu dull o greu ffynhonnell

werthfawr o ddata, lle mae'r defnyddiwr yn cyfrannu data wrth dderbyn gwasanaeth penodol.

Lemateiddiwr

Lemateiddiwr oedd y cam nesaf yn natblygiad technoleg lleferydd Cymraeg. Mae lemateiddiwr yn gallu adnabod pob ffurf o air wrth ddarganfod y lema (*lemma*) gwreiddiol. Yn ieithyddol, mae lemateiddio yn cyfeirio at y broses o grwpio ffurfiau rhediadol gair gyda'i gilydd fel bod modd eu dadansoddi fel un eitem. Yn gyfrifiadurol, mae'n broses algorithmig sy'n astudio'r geiriau i ddarganfod y lema, ac felly ystyr y gair. Yn wahanol i foniwr (*stemmer*), sy'n broses o dynnu terfyn gair i ffwrdd i ddarganfod ei wraidd, mae lemateiddiwr yn defnyddio tagio rhannau ymadrodd (*part of speech tagging*) er mwyn adnabod y lema. Mae'r cysylltiadau'n gallu bod yn gymhleth ac yn anodd mapio mewn iaith forffolegol gyfoethog fel y Gymraeg, hynny yw iaith sydd â mwy o ffurfiau ffurfdroadau (cyflyrau, rhediadau, treigladau ac yn y blaen) nag ieithoedd cyflynol (*agglutinative*), fel Saesneg i raddau. Ni fyddai boniwr yn dadansoddi'r ystyr yn gywir bob tro felly, lle byddai lemateiddiwr yn fwy tebygol o lwyddo. Mae *ellir* > *gellir* > ***gallu*** er enghraifft, yn dangos rhediad fyddai'n debygol o gael ei gamddehongli gan foniwr.

Cyhoeddwyd lemateiddiwr Cymraeg yn 2015 (Jones et al., 2015) fel gwasanaeth i allu darganfod lema gair o'i ffurf dreigledig, rediadol neu ffurfdroedig, a chwilio'r ystyr cywir. Gan fod technoleg lleferydd yn bennaf wedi'i ddatblygu yn Saesneg, mae'n codi'r cwestiwn a fyddai'r systemau hyn yn ddigon soffistigedig i ymdopi gyda gwahanol ffurfiau ieithoedd mwy cyfoethog fel y Gymraeg. Mae Meyer (2019, t. 15) yn dod i'r canlyniad na fyddai system oedd wedi'i ddatblygu ar gyfer Saesneg yn gweithio'n effeithiol ar gyfer iaith forffolegol gyfoethog gan na fyddai'r seiniau'n ymddangos yn y data hyfforddi. Mae Bohnet et al. (2013, t. 415) yn adleisio hyn, gan nodi bod ieithoedd eraill sy'n forffolegol gyfoethog yn annhebygol o weld llwyddiant gyda system sydd wedi'i ddatblygu ar gyfer Saesneg. Dengys ymchwil Avramidis & Koehn (2008) bod defnyddio Saesneg fel sail ar gyfer tasg cyfieithu peirianyddol i iaith forffolegol gyfoethog yn un cymhleth hefyd.

2.3. GALLU, Paldaruo a Macsen

Parhaodd y gwaith felly o greu adnoddau yn benodol ar gyfer ieithoedd llai eu hadnoddau yn hytrach na fforchio oddi ar adnoddau Saesneg oedd ar gael ar y farchnad. Ceir trosolwg o'r gwaith a gyflawnwyd ar gyfer y Gymraeg rhwng 2013 a 2019 yn y rhan hon.

Fel nodir yn Ffigwr 3 yn y bennod flaenorol, II 2 Adnabod Lleferydd, mae mathau gwahanol o dechnoleg adnabod lleferydd i gael:

llais-benodol:	ymateb i un llais yn unig
llais annibynnol:	ymateb i fwy nag un llais
geirfa benodol:	ymateb i eirfa gyfyngedig
geirfa eang:	ymateb i eirfa eang

Isod yn Ffigwr 10, ceir trosolwg o'r gwaith ymchwil perthnasol. Gellid rhannu'r gwaith yma i bum project (0-4, yn rhannol wedi'u tynnu o Prys & Jones (2018)), i gyd yn gweithio tuag at y nod o ddatblygu adnabod lleferydd effeithiol ar gyfer y Gymraeg.

Project 0: *prawf cysyniad (2013-2014)*

- Golygu LTS Cymraeg;
- Datblygu ap iOS ac Android ar gyfer torfoli data sain Cymraeg (lleisiau Voxforge.org oedd yr unig opsiwn cyn hyn): ap Paldauro;
- prawf cysyniad: defnyddio corpws lleferydd Paldaruo (casglwyd drwy'r ap) er mwyn adeiladu modelau acwstig i reoli braich robotig drwy cyfres o bromptiau syml a chyfyngiedig, e.e. 'braich i fyny' (Cooper et al. 2014).

Project 1: HTK & julius-cy (2015-2016)

- Wrth i gorpws lleferydd Paldaruo dyfu mewn maint drwy gasglu data o'r ap Paldaruo, symud ymlaen at y prototeip cyntaf o gynorthwyydd personol deallus: Macsen;
- Defnyddio cit offer HTK (V 1.2 Cymharu HTK a Kaldi) a dadgodiwr julius-cy er mwyn rhedeg Macsen ar *Raspberry Pi* (HTK dim ond yn gallu adeiladu modelau acwstig, roedd angen y dadgodiwr julius-cy er mwyn ei gysylltu â'r modelau iaith, *Raspberry Pi* oedd y cyfrifiadur syml cost-isel i rhedeg y feddalwedd);
- Adnabod lleferydd cyfres o gwestiynau a gorchmynion syml a chyfyngiedig e.e. 'Beth yw'r tywydd heddiw?' (Jones & Cooper 2016).

Project 2: Kaldi a *MaryTTS* (2016-2017)

- Kaldi (V 1.2 Cymharu HTK a *Kaldi*) yn gallu creu a chysylltu modelau acwstig a modelau iaith. Oherwydd hyn a'r ffaith ei fod yn denu sylw yn y gymuned technolegol, symudodd y project ymlaen i ddefnyddio *Kaldi* fel cit offer yn hytrach nac HTK. Doedd dim angen y dadgodiwr julius-cy rhagor o'r herwydd.
- Wedi bod yn defnyddio llais TIL *Ivona* hyd yn hyn (Amazon *Polly* erbyn hyn), nad oedd yn god agored. *MaryTTS* yn lais synthetig oedd yn gweithio gyda'r Gymraeg ac yn god agored felly symudodd y project ymlaen i ddefnyddio'r llais *MaryTTS.*

Project 3: cwestiynau pen-agored (2017-2018)

- Ehangu parth Macsen i gynnwys cwestiynau pen-agored yn hytrach na'r cwestiynau/gorchmynion syml cyfyngedig o Broject 1;
- Defnyddio cwestiynau Wicipedia er mwyn ehangu'r modelau iaith a defnyddio *Common Voice* (2.4 isod) a'r ap Paldauro er mwyn ehangu'r modelau acwstig i gyd-fynd;
- Arwain at fodiwl/sgìl newydd i Macsen sy'n mynd at Wicipedia er mwyn ateb cwestiynau pen-agored e.e. 'Pwy oedd Hywel Dda?'.

Project 4: ap Macsen (2018-2019)

- Datblygu Macsen i weithio fel ap ar ffonau symudol a thabledi Android;
- Defnyddio *DeepSpeech* (2.4 isod) yn hytrach na chit offer *Kaldi* ar gyfer project 4 oherwydd gwelliannau yn symlrwydd y cod.

Ffigwr 10 Projectau adnabod lleferydd Cymraeg 2013-2019

Gwelir datblygiad y projectau Cymraeg perthnasol wedi paru mathau gwahanol o adnabod lleferydd isod.

Project 0: prawf cysyniad yn eirfa benodol ac yn gweithio orau gyda llais penodol y datblygwr yn unig
Project 1: ychwanegu geirfa ehangach a mwy o leisiau
Project 2: ychwanegu geirfa ehangach eto, a mwy eto o leisiau
Project 3: geirfa ehangach sy'n ymdopi'n well â lleisiau annibynnol
Project 4: geirfa eang sy'n ymdopi â lleisiau annibynnol

Ceir manylion y projectau yn eu tro isod.

Project 0: prawf cysyniad

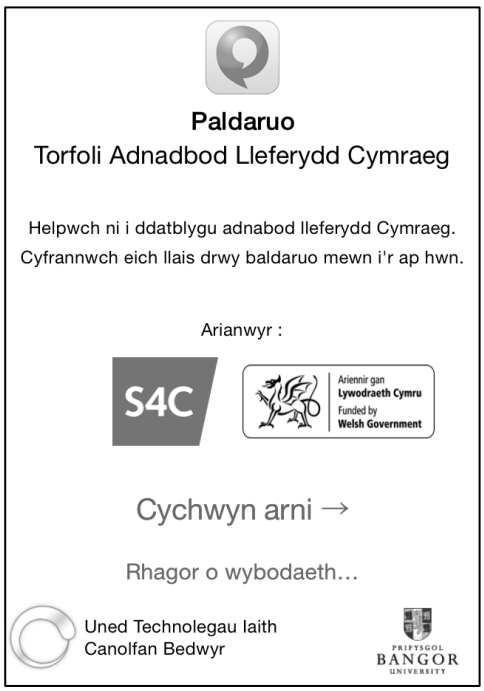
Ar ôl i strategaeth Llywodraeth Cymru, *Iaith fyw: iaith byw* (Llywodraeth Cymru, 2012), gael ei chyhoeddi yn 2012, denwyd diddordeb newydd i faes technoleg lleferydd ar gyfer y Gymraeg. Yn 2014, cyhoeddwyd y project GALLU (Gwaith Adnabod Lleferydd Uwch) (Cooper et al., 2014). Gelwir y cam yma yn Broject 0 (2013-2014) gan Prys & Jones (2018) ac adlewyrchir hyn yn Ffigwr 10 uchod. Adeiladodd y project GALLU ar waith yr Uned Technolegau Iaith yn dilyn y project synthesis lleferydd WISPR (2.2). Bwriad y project oedd gwella adnabod lleferydd er mwyn ymdopi gyda LVCSR (adnabod lleferydd geirfa eang di-dor) ar gyfer y Gymraeg (gweler II 2.1 Hanes LVCSR a Chynorthwywyr Deallus). Crëwyd set newydd o LTS (rheolau llythyren-i-sain), ond dylid nodi bod angen mwy o waith i'w gwneud yn gynhwysfawr ar gyfer trosi holl eiriau'r iaith Gymraeg (tasg enfawr). Ar hyn o bryd, dydy'r LTS ddim yn ymdopi gyda geiriau Saesneg mewn sgwrs Gymraeg (mwy ym mhennod IV 3.2 Cyfnewid Cod), er enghraifft y <v> yn '*above*'. Yn fras iawn, nid yw'r sain /v/ yn cael ei drawsgrifio o gwbl oherwydd nad oes enghreifftiau o'r llythyren 'v' yn yr LTS eto. Fel nodir ym mhennod II 2.1 Hanes LVCSR a Chynorthwywyr Deallus, mae paratoi system adnabod lleferydd ar gyfer geirfa fwy di-dor, gyda chydrannau fel LTS, yn gysyniad heriol a chymhleth oherwydd bod ffiniau annelwig rhwng ffonemau – pwynt sy'n cael ei bwysleisio gan Jones et al. (1997).
Ar ôl creadigaeth llais synthetig cod agored Cymraeg o'r project WISPR, aethpwyd ati i gasglu data newydd er mwyn ei ehangu ymhellach ar gyfer adnabod lleferydd. Crëwyd set o bromptiau seinegol gytbwys (*phonetically balanced*) i gyfranwyr adrodd, sy'n

golygu ei fod yn adlewyrchu defnydd mwyaf cyffredin seiniau'r iaith Gymraeg (Prys & Jones, 2018). Mae hon yn dechneg boblogaidd wrth gynllunio corpora sain fel gwelir ar gyfer Sbaeneg Mecsicanaidd (Uraga & Gamboa, 2004), ac i gorpws cyfnewid cod (*code switching*) Hindi-Saesneg (Pandey et al., 2018). Yn ogystal â'r promptiau newydd, ehangwyd yr LTS o'r project WISPR gan Chan & Cooper (2014), fel nodir uchod. Roedd 200 prompt (8 gair fesul prompt) oedd yn cynnwys holl ffonemau'r iaith Gymraeg; gwelir enghraifft isod o brompt fel nodir yn Cooper et al. (2014, t. 56). Byddai hyn yn arwain at greadigaeth y geiriadur ynganu sy'n cael ei ddefnyddio heddiw ar gyfer adnabod lleferydd Cymraeg. Er mwyn osgoi costau uchel a chymhlethdodau recriwtio cyfranwyr ar gyfer adnabod lleferydd, megis trefnu cyfnodau i recordio mewn bwth sain arbenigol er enghraifft, defnyddiwyd y dull torfoli (*crowdsourcing*). Gall hwn fod yn ddull cost-isel o gasglu data lleferydd fel nodir yn Cooper et al. (2019), Caines et al. (2016) a Dowling et al. (2017), er yr olaf yn y cyd-destun cyfieithu peirianyddol. Casgliad Caines et al. (2016) yw bod manteision sylweddol i'r dull torfoli, gan bwysleisio'r gost isel. Caiff hyn ei adlewyrchu yng nghasgliad Cooper et al. (2019) wrth nodi bod diffyg cyllid a ffynonellau yn heriau sylweddol ar gyfer casglu data ieithoedd llai eu hadnoddau.

Mae papur Cooper et al. (2019) yn disgrifio'r broses o gynllunio, creu a defnyddio corpws lleferydd Paldaruo (Cooper et al., 2018) ar gyfer datblygu technoleg lleferydd Cymraeg. Defnyddir y corpws hwn ar gyfer profion yn y traethawd hir hwn fel gwelir ym mhennod V 2.1 Data. Yn gwneud yn fawr o ddulliau cost effeithiol, defnyddir microffonau o ffonau symudol. Mae ansawdd y sain yn addas, a gyda chysylltiad i'r we, mae'n ddull ymarferol a chyfleus o gasglu data. O fewn yr ap, gofynnir i bob cyfrannwr greu proffil, darllen y telerau a mewnbynnu metadata am eu rhyw a'u hoedran, yn ogystal â chefndir ieithyddol a daearyddol. Gofynnir iddynt adrodd bob prompt, sy'n ymddangos fesul un.

(1) a. LLEUAD, MELYN, AELODAU, SIARAD, FFORDD, YMLAEN, CEFNOGAETH, HELEN
b. RHYBUDDIO, ELEN, UWCHRADDIO, HWNNW, BEIC, CYMRU, RHOI, AELOD

Ar ôl hynny, mae gofyn iddynt wrando ar y recordiadau yn ôl, a chadarnhau eu bod yn gywir ac o ansawdd da. Gwelir tudalen gartref yr ap Paldaruo isod yn Ffigwr 11:



Ffigwr 11 Tudalen gartref Paldaruo

Fel prawf cysyniad, defnyddiwyd y dull hwn i ddatblygu corpws bach oedd yn cynnwys geirfa i reoli braich robotig. Roedd y fraich robotig yn gallu ymateb i orchmynion penodol megis:

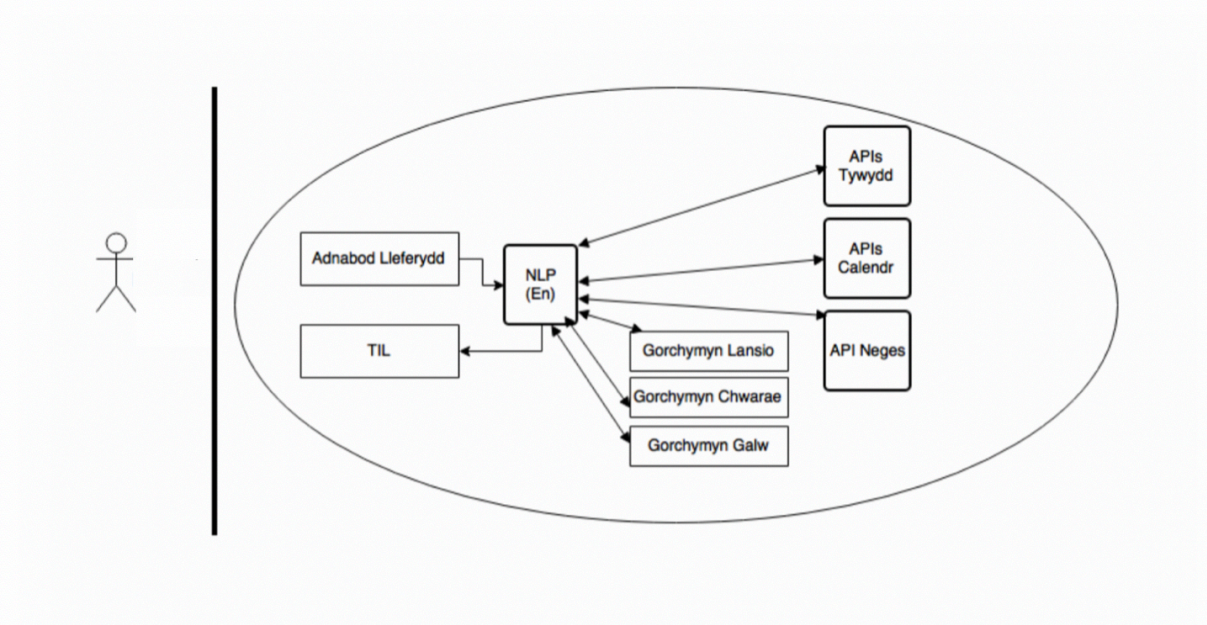
(2) a. 'braich i fyny'
b. 'braich i lawr'

Defnyddiwyd cit offer HTK er mwyn datblygu'r fraich robotig (V 1.2 Cymharu HTK a *Kaldi* (Young et al., 2015)), yn dilyn dyluniad system Saesneg oedd ar gael yn barod (AonSquared [dim cyfeiriad cyfredol]). Defnyddiwyd dadgodiwr ar wahân (*julius-cy*, isod) hefyd oherwydd nad oedd modd i HTK gysylltu modelau acwstig a modelau iaith. Roedd modd rheoli'r fraich o gyfrifiadur bach *Raspberry Pi*. Lansiwyd yr ap ar y 7fed o Orffennaf, 2014, ac fel gweddill gwaith y project GALLU, byddai'r gwaith yn bwydo mewn i greadigaeth technoleg adnabod lleferydd Cymraeg gan gynnwys gwaith yn y traethawd hir hwn.

Project 1: HTK a julius-cy

Wrth i gorpws lleferydd Paldaruo dyfu mewn maint, y cam nesaf oedd creu'r prototeip cyntaf o gynorthwyydd personol deallus. Dyma yw ffocws gweddill y projectau yn Ffigwr 10, gan ei fod yn rhan mor allweddol o ddatblygiad adnabod lleferydd Cymraeg diweddar. Cyn hyn, defnyddiwyd meddalwedd *Jasper* (http://*Jasper*project.github.io) ar gyfer adnabod lleferydd Cymraeg (cod sgiliau, cod adnabod lleferydd a'r cod testun-i-leferydd (TIL) mewn un), ond doedd *Jasper* ddim yn cydweithio â ffonau symudol. Gan nad oedd defnydd *Raspberry Pi* yn ymarferol yn yr hir dymor, a bod ffonau symudol yn blatfform gwell ar gyfer y dasg, roedd angen symud i ffwrdd o *Jasper* a thuag at fersiwn annibynnol.

Yn 2016, dechreuodd ymchwil ar Seilwaith Cyfathrebu Cymraeg yn Uned Technolegau Iaith (Jones & Ghazzali, 2016). Gellir galw'r cam hwn yn Broject 1 yn ôl Ffigwr 10 uchod. Ar ddechrau'r project Seilwaith Cyfathrebu Cymraeg, cynhelir ymchwil pellach gan Ghazzali et al. (2015) er mwyn pwyso a mesur y pedwar prif gynorthwywyr personol deallus oedd ar y farchnad ar y pryd (Apple Siri, Amazon Alexa, Microsoft Cortana a Google Now). Bwriad yr ymchwil oedd darganfod os byddai modd ehangu'r offer i iaith arall fel y Gymraeg. Cyhoeddwyd ffigwr tebyg i'r isod, Ffigwr 12, er mwyn rhoi syniad o bensaernïaeth y feddalwedd gyffredinol ar gyfer cynorthwyydd deallus.



Ffigwr 12 Pensaernïaeth system cynorthwyydd deallus yn fras

(wedi seilio ar Ghazzali et al. (2015, t. 4))

Casgliadau ymchwil Ghazzali et al. (2015) oedd nad oedd un o'r pedwar system fasnachol yn addas i'w mabwysiadu gan y Gymraeg gan eu bod yn rhy gaeedig. Yn ddelfrydol, byddai modd adeiladu ar system oedd yn bodoli yn barod ac integreiddio'r Gymraeg, felly yn creu system fyddai'n gweithio gyda chydran cyfieithu peirianyddol er enghraifft. Byddai hyn yn gallu ymdopi â mewnbwn Cymraeg, gyda chydrannau Saesneg, ond yn creu allbwn Cymraeg. Petai hyn yn bosib, gwelir y cydrannau angenrheidiol isod:

1. system barod sy'n gallu cael ei ehangu i gynnwys iaith arall fel y Gymraeg;
2. prosesydd iaith naturiol (*Natural Language Processor*, NLP) sy'n deall y Gymraeg, yn prosesu a dadansoddi'r iaith naturiol ar ffurf testun;
3. testun-i-leferydd (TIL) Cymraeg i greu'r allbwn;
4. cydran cyfieithu peirianyddol Cymraeg-Saesneg, Saesneg-Cymraeg.

Mae Jones & Cooper (2016) yn pwysleisio mai prif wendid y cysyniad hwn yw bod angen adnabod lleferydd galluog iawn sy'n gallu adnabod unrhyw gwestiwn neu orchymyn.

Yn ei le felly, defnyddir cyfuniad o'r cit offer HTK (V 1.2 Cymharu HTK a *Kaldi*) a'r dadgodiwr *julius-cy* fel sail ar gyfer creu system Gymraeg yn y project Seilwaith Cyfathrebu Cymraeg. Bwriad y rhan adnabod lleferydd o'r project oedd sicrhau nad oedd siaradwyr Cymraeg dan anfantais yn y maes technoleg adnabod lleferydd, yn benodol, wrth ddefnyddio cynorthwywyr digidol (Jones & Cooper, 2016). Fel gweddill gwaith yr Uned Technolegau Iaith yn y maes hwn, byddai'r gwaith yn cael ei gyhoeddi dan drwydded agored fel bod ieithoedd eraill yn gallu manteisio ar y fethodoleg hefyd (1.2). Yn hytrach na dim ond defnyddio is-set o ddata Paldaruo fel a wnaed yn y project GALLU (Cooper et al., 2014), roedd y project hwn yn defnyddio'r corpws cyfan, oedd wedi tyfu mewn maint ers 2014 i gynnwys 410 cyfrannwr. Golygwyd ac ehangwyd y sgriptiau hyfforddi o'r project GALLU i gymhwyso'r anghenion newydd, er enghraifft llwytho'r corpws i lawr. Yn anffodus, oherwydd nad bwriad gwreiddiol data Paldaruo oedd bwydo fewn i gynorthwyydd personol deallus, doedd dim set brofi ar gael, dim ond set hyfforddi (rhaniad 10% a 90% o'r data, fesul set, fel arfer, fel nodir ym mhennod II 4 Dulliau i Ymdopi gyda Llai o Ddata). Oherwydd hyn, roedd rhaid dilyn dull arfer gwael (ond angenrheidiol ar y pryd) o ddefnyddio rhan o'r data hyfforddi fel data profi. Gyda modelau acwstig wedi hyfforddi yn y modd hwn, aethpwyd ati ym Mhroject 1, Ffigwr 10,

i ddatblygu prototeip cynorthwyydd digidol Cymraeg. 'Macsen' oedd y persona a'r gair deffro, a fyddai'n ei drosi o fod mewn modd segur i fod yn weithredol. Roedd modd i Macsen ymateb i gwestiynau a gorchmynion parth-benodol ym meysydd newyddion, tywydd, amser, cerddoriaeth, ymadroddion a jôcs. Gwelir isod rhai enghreifftiau o'r promptiau.

(3) a. 'Beth ydy'r tywydd heddiw?';
b. 'Beth yw tywydd yfory?';
c. 'Beth yw'r newyddion?';
ch. 'Faint o'r gloch ydy hi?';
d. 'Chwaraea gerddoriaeth Cymraeg.'.

Daw llais Macsen o'r llais a ddatblygwyd o *Festival* (Black & Lenzo, 2000), ac o 'Geraint' gafodd ei ddatblygu gyda Ivona (Amazon Polly erbyn hyn, https://aws.amazon.com/polly/).

Project 2: Kaldi a Mary TTS

Erbyn 2016-2017, roedd y cit offer *Kaldi* (V 1.2 Cymharu HTK a *Kaldi* (Povey, et al., 2011)) yn cael ei ddefnyddio'n aml o fewn y gymuned dechnolegol. Oherwydd bod *Kaldi* yn gallu cysylltu modelau acwstig a modelau iaith heb yr angen am ddadgodiwr fel *julius-cy*, defnyddiwyd cit offer *Kaldi* yn lle HTK i ddatblygu Macsen o hynny ymlaen. Gan nad oedd llais Ivona (uchod) yn god agored, roedd datblygiad llais *MaryTTS* (http://mary.dfki.de), oedd yn god agored, yn ychwanegiad yn 2016.

Project 3: cwestiynau pen-agored

Y cam nesaf tuag at gynorthwyydd deallus oedd yn gallu ymdopi â lleferydd geirfa ehangach oedd cwestiynau pen-agored. Ym Mhroject 1 a 2, roedd Macsen yn gallu ymateb i gasgliad o gwestiynau neu orchmynion penodol a chyfyngedig iawn (gwelir enghraifft uchod). Er mwyn symud ymlaen at gwestiynau pen-agored, defnyddir Wicipedia i gasglu mwy o ddata (Prys & Jones, 2018). Crëwyd modelau iaith newydd (cyflwyniad i'r modelau ym mhennod II 3.1 Modelau Ystadegol) o gwestiynau mwyaf cyffredin Wicipedia. Mewnbynnwyd is-set o'r cwestiynau pen-agored hyn i'r ap Paldaruo (Cooper et al., 2014). Erbyn hyn, maent wedi'u mewnbynnu i blatfform *CommonVoice-cy* hefyd, mwy isod yn 2.4. Dengys Jones et al. (2019, t. 28) bod y dull hwn o ddefnyddio

brawddegau Wicipedia yn llwyddiannus ond yn tynnu sylw at rhai gwendidau yn y corpws wrth ymateb i gwestiynau pen-agored. Gwelir cymhariaeth o ganlyniadau cwestiynau cyfyngedig, parth-benodol Macsen, isod (Tabl 2) yn erbyn canlyniadau cwestiynau pen-agored Wicipedia sy'n dangos y gwendid hwn. Defnyddir y metrig WER sy'n cyfrif nifer y gwallau felly yr isaf yw'r rhif, y gorau oll yw'r canlyniad (manylion llawn ym mhennod V 2.2 Gweithdrefn).

Macsen: parth-benodol	**Macsen: cwestiynau pen-agored**
2.24% WER	65.88% WER

Tabl 2 Cymharu canlyniadau Macsen: parth-benodol, a Macsen: cwestiynau pen-agored

Project 4: ap Macsen

Yn ystod 2018–2019, buodd yr Uned Technolegau Iaith yn datblygu Macsen (Techiaith, 2019) i weithio fel ap ar ffonau symudol a thabledi Android. Roedd yr ap yn defnyddio technoleg lleferydd-i-destun er mwyn trosi'r lleferydd fewn i destun, yna yn defnyddio parsio bwriad (*intent parsing*) er mwyn darganfod bwriad a chaffael y data cywir (newyddion, tywydd, cerddoriaeth ac yn y blaen), ac yna TIL er mwyn llefaru'r ymateb priodol. Erbyn mis Ebrill 2019 roedd chwe sgìl gan yr ap Macsen:

1. gosod larwm;
2. dweud yr amser;
3. rhoi'r dyddiad;
4. darllen y newyddion (o Golwg 360, https://golwg.360.cymru);
5. adrodd am y tywydd (o OpenWeatherMap, https://openweathermap.org);
6. chwarae cerddoriaeth (o Spotify, https://www.spotify.com/uk/).

Wrth ofyn am y newyddion, roedd modd gofyn am benawdau pwnc arbennig, newyddion Cymru yn benodol neu chwaraeon er enghraifft. Wrth ofyn am gerddoriaeth, roedd yn adnabod 21 band ac unigolyn Cymraeg; rhywbeth nad ydy cynorthwywyr masnachol fel Alexa, Siri a Google Home yn medru gwneud yn effeithiol ar y foment (rhestr gyfan ar Techiaith (2019)). Fel nodir uchod yn 2.1, mae'r LTS yn cael effaith uniongyrchol ar allu

system i adnabod seiniau penodol, megis /ɬ/ ar ddechrau'r enw 'Lleuwen'. Gan fod y sain yma yn ymddangos yn LTS Macsen, ond yn annhebygol o ymddangos yn LTS Saesneg cwmnïau masnachol (nid yw'r data y defnyddia cwmmïau masnachol yn gyhoeddus, felly nid os modd dweud o sicrwydd), mae gwendid wrth adnabod enwau megis 'Lleuwen'.

Dyma felly yw un o brif amcanion creu cynorthwyydd personol deallus Cymraeg, er mwyn sicrhau bod yr iaith Gymraeg yn gallu cael ei defnyddio o fewn maes poblogaidd adnabod lleferydd, rhywbeth sydd ddim o reidrwydd yn bosib wrth ddefnyddio'r systemau sydd ar y farchnad yn barod. Gwelir yng ngwaith timau ymchwil eraill llai eu hadnoddau megis Sindhi (Jamro, 2017), Arabeg (Khelifa et al., 2017) a Ffrisieg (Yılmaz, et al., 2016), fel gwelir uchod yn 1.1, mai'r opsiwn mwyaf ymarferol yn aml yw mynd ati i gasglu data newydd.

2.4. Mozilla, *Common Voice* a *Common Voice* Cymraeg

Mae Mozilla yn gefnogwyr brwd o ieithoedd llai eu hadnoddau yn y maes erbyn hyn. Mozilla yw'r cwmni mwyaf sy'n darparu offer a deunyddiau yn agored mewn cyd-destun amlieithog i drawstoriad eang o ieithoedd, gan gynnwys ieithoedd llai eu hadnoddau. Mae'n gwmni nid-er-elw sydd y tu ôl i'r peiriant chwilio *Firefox.* Eu prif nod yw galluogi pawb i ddefnyddio'r we. Maent yn canolbwyntio felly ar greu a hyrwyddo deunyddiau a phlatfformau agored ar y we, ac wedi troi at adnabod lleferydd yn ddiweddar. Ym mis Gorffennaf 2017, cyhoeddwyd project adnabod lleferydd Mozilla: *Common Voice* (https://commonvoice.mozilla.org) (Ardila, et al., 2020, t. 4218). Bwriad y project oedd sicrhau bod adnabod lleferydd yn agored i unrhyw un drwy gasglu data newydd. Datblygwyd y platfform ar ap a gwefan gyfatebol gan ddefnyddio'r dull torfoli, yn debyg i Paldaruo. Ar yr ap *Common Voice*, mae defnyddwyr yn gallu cyfrannu eu lleisiau wrth adrodd promptiau a'u gwirio ond, yn wahanol i Paldaruo, mae modd dilysu cyfraniadau gan bobl eraill hefyd. Mae'r gwirio yn digwydd drwy system ddilysu lle mae dau neu dri gwirfoddolwr yn marcio recordiad yn gywir neu'r anghywir, gyda dau farciwr yn gorfod cytuno fod y recordiad yn gywir cyn y caiff ei dderbyn – dull effeithiol o wirio ansawdd y corpws.

Ar y 7fed o Fehefin 2018, cyhoeddodd Mozilla ehangiad i'r platfform *Common Voice* i ieithoedd eraill, gan gynnwys y Gymraeg (Ardila, et al., 2020). Gyda chymorth yr Uned Technolegau Iaith, fe grëwyd *Common Voice-cy* (https://voice.mozilla.org/cy). Er mwyn ceisio creu bwrlwm ac i dynnu sylw, cyhoeddwyd her i gyrraedd 100 awr o recordiadau. Cyfrannodd sefydliadau megis Cynulliad Cenedlaethol Cymru (bellach

Senedd Cymru) (Prys R., 2019) a Mentrau Iaith Cymru (2019) eu lleisiau i helpu *Common Voice-cy* gyrraedd y nod hwn. Blwyddyn ar ôl cyhoeddiad yr ymgyrch *Common Voice-cy*, roedd 42 awr wedi recordio a 28 awr wedi eu dilysu felly cynnydd addawol, ond fel noda R. Prys mewn erthygl ar-lein, bod 'tipyn o ffordd i fynd' (2019). Gydag ymdrechion ychwanegol yn Eisteddfod Genedlaethol 2019 (Prys R., 2019), tyfodd hwn i 77 awr wedi recordio a 59 awr o'r rheiny wedi'u dilysu (1,149 cyfrannwr) erbyn y cyhoeddiad data cyntaf ar y 10fed o Ragfyr, 2019. Gwelir y twf mewn maint data isod yn Tabl 3:

Dyddiad	**Oriau**	**Dilysu**	**Cyfranwyr**
07/06/2017	0	0	0
31/08/2017*	5	0	94
11/10/18**	18	16	249
21/06/19***	52	44	762
10/12/19****	77	59	1,149

* Cofnod cyntaf o gyfraniadau

** Cofnod cyntaf o gyfraniadau wedi'u dilysu

*** Cefnogaeth mentrau a sefydliadau Cymreig

**** Cyhoeddiad data cyhoeddus cyntaf

Tabl 3 Ystadegau *Common Voice-cy*

(gyda diolch i Rhoslyn Prys am y data)

Mae *Common Voice-cy* yn cynnig platfform newydd ar gyfer y dull torfoli (uchod), sy'n addas ar gyfer casglu data iaith lai ei hadnoddau. Erbyn 2019, roedd data *Common Voice-cy* yn cael ei ddefnyddio ar gyfer hyfforddi a phrofi system adnabod lleferydd Cymraeg ac mae'r traethawd hir hwn yn cynnal profion ym mhennod VI Math o Ddata, er mwyn arbrofi i weld os yw'r data yn rhoi canlyniadau gwell na data Paldaruo, ac os felly pam bod hynny'n wir.

3. Casgliadau

Nod y bennod hon oedd cynnig trosolwg o'r gwaith sydd wedi'i gyflawni a'r adnoddau sydd wedi'u creu ar gyfer technoleg lleferydd Cymraeg o ganlyniad i ddiffyg adnoddau yn y maes. Noda Berger et al. (2018) bod datblygu adnoddau iaith yn hanfodol ar gyfer adferiad yr iaith yn ddigidol, ac o ddeall rhai o'r heriau sy'n wynebu ieithoedd llai eu hadnoddau fel gwelwyd yn 1.1, mae timau ymchwil wedi mynd ati i greu'r adnoddau angenrheidiol er mwyn datblygu technoleg lleferydd yn eu hieithoedd eu hunain.

Cafwyd cyflwyniad i ddechreuadau casglu data a'r project SpeechDat Cymru yn 2.1, ac i WISPR a phrojectau diweddarach yr Uned Technolegau Iaith yn 2.2, 2.3 a 2.4, lle cyflawnir gwaith allweddol yn y maes. Cyflwynwyd cysyniadau LTS (rheolau llythyren-i-sain) a lemateiddiwr yn yr is-rannau hyn hefyd, sy'n dal i gael eu defnyddio mewn adnabod lleferydd cyfredol. Cafwyd cyflwyniad i waith GALLU, Paldaruo a Macsen yn benodol yn 2.3, sy'n dangos y broses o ddatblygu cynorthwyydd personol deallus Cymraeg. Arweiniodd hyn oll at lwyddiant *Common Voice-cy*, sy'n bwydo mewn i'r datblygiad mwyaf diweddar, *DeepSpeech* (VIII *DeepSpeech*). Y cam nesaf yw asesu sut orau i barhau â'r gwaith yn wyneb yr heriau sydd wedi'u disgrifio yn y bennod hon a'r nesaf. Arweinia hyn at ymchwilio i fathau gwahanol o ddata a sut orau i ymdopi gyda'r heriau cysylltiedig, er mwyn adeiladu system gadarn yn y penodau dull i ddilyn. Gwelir yr amcanion a chwestiynau ymchwil sy'n arwain ymlaen o hyn oll ar ddiwedd y bennod nesaf, IV 5 Amcanion yr Ymchwil. Mae'r bennod nesaf, IV Cefndir: Amrywio yn y Gymraeg, yn edrych yn fanylach ar sut gall amrywio iaith fod yn her prosesu ychwanegol i'r broses o adnabod lleferydd.

IV. Cefndir: Amrywio yn y Gymraeg

1. Cyflwyniad

Pwrpas y bennod hon yw adeiladu ar y cyflwyniad a geir yn y bennod flaenorol i ieithoedd llai eu hadnoddau a thechnoleg. Yn y bennod honno, tynnwyd sylw at y diffyg data sy'n gyffredin ymysg ieithoedd llai eu hadnoddau, a'r gwaith sydd wedi'i gyflawni o'r herwydd. Fel nodir ym mhennod I 2.2 'Iaith Lai ei Hadnoddau', ac eto ym mhennod III 2.2 WISPR, Cysill Ar-lein a Lemateiddiwr, mae'r rhan fwyaf o ddatblygiadau technoleg lleferydd wedi'u datblygu ar gyfer ieithoedd mwy eu hadnoddau fel Saesneg. Hwn yw un o'r prif ffactorau sy'n cyfrannu at ddiffyg adnoddau fel data, ar gyfer ieithoedd llai eu hadnoddau fel y Gymraeg. Lle bo data yn barod yn brin, mae ffactorau ychwanegol sy'n gallu ychwanegu at yr her. Ceir yn y bennod hon gyflwyniad i ffurfiau gwahanol yr iaith Gymraeg yn benodol, i'r nifer o ffurfiau gwahanol sy'n cael eu cynhyrchu ar sail ei morffoleg gyfoethog ac i amrywiadau eraill rhwng siaradwyr hefyd. Gall hyn oll ychwanegu at yr angen am ddata er mwyn cynnwys yr holl ffurfiau, ac felly at yr her prosesu ar gyfer adnabod lleferydd.

Cyfeirir y term 'morffoleg gyfoethog' at iaith sydd â nifer fawr o ffurfdroadau gwahanol ar gyfer y lemma gwreiddiol (Huck et al., 2017). Mae ieithoedd morffolegol gyfoethog fel y Gymraeg, sydd â nifer fawr o ffurfdroadau fel gwelir yn 2. Mae hyn yn aml yn golygu bod ieithoedd fel y Gymraeg yn y modd hwn yn dioddef o ddiffyg data beth bynnag, oherwydd eu bod, wrth reswm, yn cynhyrchu nifer fawr o ffurfiau gwahanol, ac felly yn gofyn am feintiau mwy o ddata i allu cynnwys yr holl ffurfiau (Niehues, et al., 2018; Buys & Botha, 2016). Mae'r diffyg data hwn yn gallu arwain at nifer uchel o eiriau yn y rhestr allan o'r eirfa (*out of vocabulary*, OOV), sy'n arwain at nifer uwch o wallau (canlyniadau WER uwch), ac at fodelau gwannach yn y pendraw (Sak et al., 2010). Mewn sefyllfa iaith lai ei hadnoddau sy'n barod yn debygol o ddioddef o ddiffyg data, mae'r geirffurfiau niferus hyn yn ychwanegu at yr her prosesu. Ceir cyflwyniad i rai enghreifftiau o sut mae'r Gymraeg yn cynhyrchu ffurfiau gwahanol sy'n berthnasol i'r gwaith hwn yn 2. Ceir hefyd enghreifftiau o amrywiadau gwahanol fel cywair, cyfnewid cod, tafodiaith ac acen, sydd ddim ar sail morffoleg yn 3. Gall yr olaf fod yn her ychwanegol i adnabod lleferydd, am eu bod yn achosi ffurfiau gwahanol eto. Mae'r amrywiadau hyn yn gallu codi oherwydd cywair neu dafodiaith wahanol er enghraifft, all amrywio rhwng siaradwyr (Nordquist, 2020). Oherwydd bod cymaint o

eirffurfiau gwahanol yn bosib, mae gofyn am ddata sylweddol i hyfforddi'r modelau iaith, a lle nad oes mwy o ddata ar gael fel ar gyfer ieithoedd llai eu hadnoddau, mae hyn yn gallu arwain at fodelau gwannach.

Ceir cyflwyniad i gyd-destun ehangach ieithoedd llai eu hadnoddau eraill yn 1.1. Mae'r bennod wedi rhannu i ddwy ran: cyflwyniad i eirffurfiau yn y Gymraeg yn 2, a chyflwyniad i amrywiadau yn y Gymraeg yn 3, ill dau yng nghyd-destun eu bod yn her ychwanegol i brosesu. Gwelir y cwestiynau ymchwil a'r amcanion sy'n arwain ymlaen o hyn, ac o'r penodau blaenorol, yn 5.

1.1. Ieithoedd Llai eu Hadnoddau ac Amrywio

Diffinnir iaith lai ei hadnoddau (I 2.2 'Iaith Lai ei Hadnoddau') fel iaith lle mae diffyg rhai agweddau megis adnoddau digidol ar gyfer prosesu iaith a lleferydd. Nododd Besacier et al. (2014, t. 86), yn 2014, bod mwy na 6,900 iaith yn y byd yn dioddef o ddiffyg adnoddau ar gyfer technoleg iaith, gyda lan at 21 iaith (ac acenion Saesneg gwahanol) yn cael eu cefnogi gan Apple Siri yn 2020 er enghraifft (Templeton, 2020). Mae ieithoedd llai eu hadnoddau felly yn dioddef o anfantais wrth ddatblygu technoleg iaith megis adnabod lleferydd. Wrth ystyried amrywio yn yr ieithoedd hyn, nodir uchod bod angen mwy o ddata i ymdopi ag amrywio er enghraifft, sy'n her ychwanegol mewn sefyllfa sy'n barod yn heriol am y rhesymau uchod. Mae gwaith Tarján (2019) er enghraifft, yn edrych ar dechnoleg iaith Hwngareg. Mae'r iaith yn forffolegol gyfoethog, all feddwl bod diffyg data i gynnwys yr holl ffurfiau. Mae amrywiadau cywair hefyd yn her i Tarján (2019) yn ei waith oherwydd natur anffurfiol y data dan sylw. Mae'r gwaith yn ffocysu ar ddata dros y ffôn mewn arddull anffurfiol, sgyrsiol, lle mae defnydd gramadeg a threfn y frawddeg yn amrywio, sy'n gwneud y modelau yn llai effeithiol wrth brosesu'r lleferydd gan fod y posibiliadau o ddilyniant fesul gair yn amrywio (2019, tt. 1-2). Gall hyn wedyn arwain at ganlyniadau gwaeth. Caiff yr union heriau hyn eu hadleisio mewn gweithiau ymchwil eraill ar draws y maes llai eu hadnoddau, ar gyfer iaith Uyghur (Abulimiti & Schultz, 2020), Amharic (Tachbelie et al., 2009) a Thwrceg (Sak, Saraşlar, & Güngör, 2009; 2010) er enghraifft. Mae'r heriau o fod yn iaith lai ei hadnoddau yn y maes technoleg lleferydd yn anoddach fyth felly wrth fod yn iaith sydd yn meddu ar lawer o amrywio ieithyddol. Er bod pob iaith yn dangos rhywfaint o amrywio, mae'r sefyllfa yn fwy heriol ar gyfer ieithoedd llai eu hadnoddau gan eu bod yn dioddef o ddiffyg data hefyd, a bod angen mwy o ddata i gynnwys yr holl ffurfiau hyn. Heb ddata

digonol, does dim digon o enghreifftiau o'r amrywio, sy'n cyfrannu at nifer uwch o eiriau OOV, sy'n cynrychioli nifer uwch o wallau, fel nodir uchod.

Mae'r rhannau nesaf yn cyflwyno heriau amrywio ar gyfer y Gymraeg yn benodol sy'n berthnasol i'r traethawd hir hwn. Rhennir y mathau o amrywio i eirffurfiau sy'n cael eu cynhyrchu gan y Gymraeg ar sail ei morffoleg, yn 2, ac yna amrywiadau gwahanol sy'n deillio o amrywio cywair a thafodiaith er enghraifft, yn 3. Maent ill dau yn gallu cyfrannu at nifer uwch o wallau, lle nad yw'r ffurfiau gwahanol wedi'u cynnwys yn y data hyfforddi ac felly ddim yn cael eu hadnabod gan y modelau. Ar gyfer y Gymraeg, sy'n iaith lai ei hadnoddau ac yn dioddef o ddiffyg data yn barod, mae hwn yn ychwanegu at yr her prosesu.

2. Geirffurfiau yn y Gymraeg

Mae'r traethawd hir hwn yn cynnal profion yn y penodau i ddilyn ar fathau gwahanol o ddata Cymraeg ar gyfer adnabod lleferydd, ac felly ceir yma rhai enghreifftiau o ffurfiau gwahanol all gael eu cynhyrchu o forffoleg yr iaith ei hun. Gwneir hyn er mwyn rhoi cipolwg ar y fath o heriau prosesu all godi wrth adnabod lleferydd Cymraeg o'r iaith, cyn ystyried amrywiadau rhwng siaradwyr hefyd yn 3. Wrth gyflwyno'r ffurfiau gwahanol yn 2, cyfeirir at sut gall y ffurfiau newid eto yn dibynnu ar y siaradwr ac ar gyd-destun y sgwrs; caiff hwn ei ystyried yn fanylach yn 3 fel nodir uchod.

Fel nodir yn 1, mae'r Gymraeg yn iaith gynhyrchiol sy'n creu nifer o eirffurfiau gwahanol ar sail ei morffoleg gyfoethog. Mae ieithoedd morffolegol gyfoethog yn gofyn am ddata mwy er mwyn cynnwys yr holl ffurfiau gwahanol. Yn aml, mae'r ieithoedd yn cynhyrchu nifer uwch o eiriau OOV (allan o'r eirfa) oherwydd diffyg data digonol, ac mae effeithiolrwydd y modelau yn dirywio o ganlyniad (Niehues, et al., 2018, t. 6). Mae Pellegrini & Lamel (2009, t. 865) yn amcangyfrif bod iaith forffolegol gyfoethog yn debygol o greu 1.5 i 2 gwall am bob gair yn yr OOV, sy'n cael effaith negyddol ar fodelau'r system a'i effeithiolrwydd o ganlyniad.

Isod, ffocysir ar ffurfiau gwahanol sy'n cael eu cynhyrchu gan y system dreiglo, y ferf afreolaidd 'bod', arddodiaid rhediadol, a threfn geiriau o fewn brawddeg y Gymraeg. Gwneir hyn am eu bod yn rhai agweddau anghyffredin neu wahanol mewn ieithoedd eraill y byd, ac am eu bod yn cynhyrchu nifer o ffurfiau gwahanol all fod yn her prosesu.

2.1. Treiglo

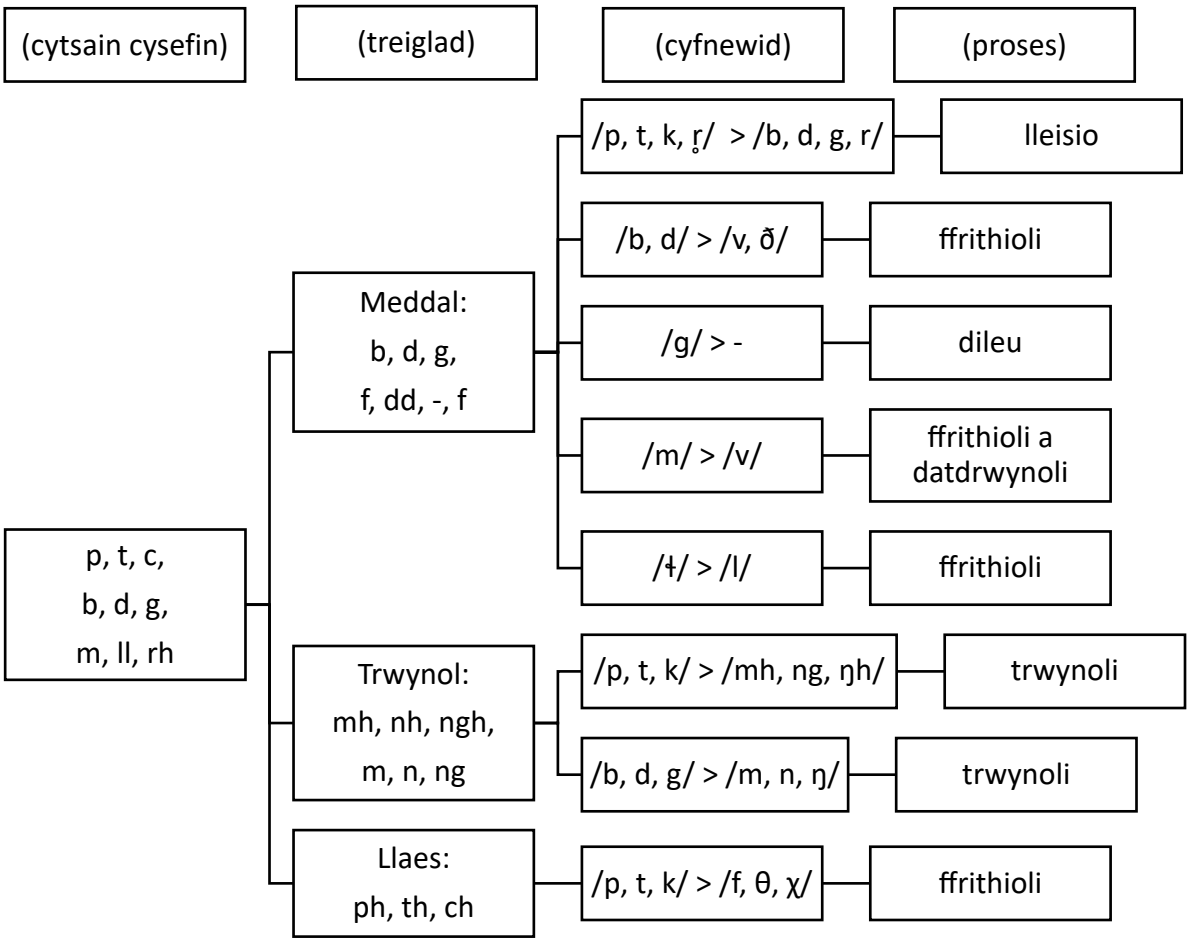
Mae'r system dreiglo yn cynhyrchu nifer o ffurfiau gwahanol o'r un gair, yn dibynnu ar gyd-destun gweddill y frawddeg, fel esbonnir isod. Gall hwn fod yn her prosesu am fod angen modelu ar gyfer pob ffurf dreigledig, sy'n gofyn am fwy o ddata er mwyn prosesu.

Gan nad oes consensws clir ymysg ymchwilwyr ar gategori ieithyddol treigladau, mae Borsley et al. (2007, tt. 223-254) yn nodi bod treigladau yn fath o eilededdau morffoffonolegol sy'n achosi amrywio mewn cytseiniaid rhai geiriau. Er bod ffurfiau treigledig yn cael eu cynhyrchu o strwythur morffolegol yr iaith, dylid nodi bod modd amrywio defnydd y treiglo yn ôl arddull y siaradwr a chyd-destun y sgwrs hefyd (cyflwyniad i hyn yn 3.1). Er bod rhai, megis Oftedal (1985), wedi awgrymu bod treiglo yn cael ei rhannu gan ieithoedd eraill megis Sbaeneg, ac ieithoedd eraill Lladinaidd yn ôl Martinet (1952), y farn gyffredinol yw ei bod yn nodwedd unigryw i'r Gymraeg a'r ieithoedd Celtaidd (Ball & Müller, 1992), (Fife, 2002, tt. 8-9).

Mae treigladau yn effeithio cytsain gyntaf rhai geiriau mewn rhai amgylcheddau cystrawennol a morffolegol. Mae tri math gwahanol o dreiglad yn y Gymraeg: y treiglad meddal, y treiglad trwynol, a'r treiglad llaes. Gwelir isod effaith y treigladau ar yr enw benywaidd 'cath':

(4) a. cath (ffurf gysefin)
b. *ei **g**ath* (meddal)
c. *fy **ngh**ath* (trwynol)
ch. *ei **ch**ath* (llaes)

Gwelir isod yn Ffigwr 13, y newidiadau sy'n rhan o dreiglo. Mae'n dangos y newidiadau sy'n digwydd ar gyfer y tri math o dreiglo: meddal, trwynol, a llaes. Cawsant eu sbarduno mewn gwahanol gyd-destunau megis meddal mewn enw benywaidd ar ôl y fannod (uchod), trwynol mewn enwau, ac eraill, yn dilyn yr arddodiad 'yn', a llaes mewn enwau yn dilyn rhai rhifau megis 'tri' neu 'chwe'. Mae eu prosesau yn wahanol hefyd fel gwelir isod yn Ffigwr 13: lleisio, ffrithioli, dileu, neu drwynoli.



Ffigwr 13 Prosesau treiglo

(wedi seilio ar (Thomas P. W., 1996))

Dylid nodi rhagddodi ‘h’ hefyd wrth drafod treiglo. Mae’r newid yma yn dal i fod yn un ffonolegol sy’n cael ei effeithio gan amgylcheddau cystrawennol, ond yn hytrach na dileu neu newid cytsain gychwynnol gair, mae’n rhagddodi /h/ o flaen llafariad yn y cyd-destunau isod.

(5) a. ar ôl rhai rhagenwau, ‘ *‘m*, *ei* (b), *‘i*, *‘w* (b), *ein*, *‘n*, *eu*, *‘u* ’
b. ar ôl yr arddodiad ‘*ar*’

Fel nodir uchod, mae treiglo yn un o’r heriau ieithyddol sy’n gallu peri trafferth i brosesu lleferydd Cymraeg gan ei fod yn achosi ffurfiau gwahanol, sy’n ychwanegu mwy o waith prosesu cyfrifiadurol.

2.2. Arddodiaid Rhediadol

Fel y system dreiglo, dydy arddodiaid rhediadol ddim yn gyffredin ymysg ieithoedd y byd, ond maent yn codi'n aml iawn yn y Gymraeg. Mae arddodiaid yn cael eu defnyddio fel ymateb i enw neu ferfenw (Williams S. J., 1980, t. 167). Yn ogystal â bod yn sbardun ar gyfer treiglad, mae arddodiaid yn gallu bod yn heriol i brosesu'n gyfrifiadurol am eu bod yn medru rhedeg, ac felly yn newid eu terfyniadau yn ôl y person (Borsley, Tallerman, & Willis, 2007, t. 17). Mae (I) *am*, *ar*, *at*, *dan*, (II) *heb*, *rhag*, *rhwng*, *yn*, *o*, *tros*, *trwy*, *er*, *hyd* (ambell waith), (III) *gan*, *wrth* yn rhedeg fel gwelir isod (Watkins T. A., 2002, tt. 314-315):

	Unigol	**Lluosog**
1	(I) *-af*, (II) *-of*, (III) *-yf*	(I) *-om*, (II) *-om*, (III) *-ym*
2	(I) *-at*, (II) *-ot*, (III) *-yt*	(I) *-och*, (II) *-och*, (III) *-ych*
3 (g)	(I) *-o*, (II) *-ddo*/*-to*/*-o*, (III) *-o*/*-ddo*	(I) *-ynt*, (II) *-ddynt*/*-tynt*/*-ynt*, (III) *-ynt*/*-ddynt*
3 (b)	(I) *-i*, (II) *-ddi*/*-ti*/*-i* (III) *-i*/*-ddi*	

Tabl 4 Arddodiaid rhediadol

Fel gwelir uchod yn Tabl 4, mae arddodiaid yn rhedeg fesul cenedl a pherson, fel mae berfau'n gwneud, er â therfyniadau gwahanol (Borsley et al., 2007, t. 199). Dydy hyn ddim yn digwydd gydag arddodiaid yn Saesneg, lle maent yn aros yn ei strwythur gwreiddiol, yn annibynnol o amser neu berson. Gwelir enghraifft yr arddodiad 'i', '*to*', isod.

(6) 'iddo fe, iddi hi, iddyn nhw'
'*to him*, *to her*, *to them*'

Gan fod arddodiaid yn rhedeg yn y Gymraeg, mae'n rhaid i system adnabod lleferydd ymdopi gyda newidiadau yn strwythur yr arddodiad yn dibynnu ar amser y frawddeg a pherson y siaradwr. Mae hyn eto, yn ychwanegu at nifer y ffurfiau sydd angen bod yn y data, fel gyda threiglo a'r ferf 'bod' er enghraifft, yn gallu bod yn her prosesu.

2.3. Berfau Afreolaidd

Nodir uchod bod geirffurfiau gwahanol o eiriau yn gallu achosi gwaith prosesu ychwanegol gan fod angen prosesu mwy o ffurfiau o'r un gair, a pharatoi data ar gyfer yr holl ffurfiau gwahanol. Gwelir yma enghraifft o'r ferf afreolaidd 'bod'. Caiff 'bod' ei gynnwys yma fel enghraifft gan ei fod yn ymddangos yn wahanol iawn i'r gair cysefin wrth ei redeg ac ei fod yn enghraifft dda o gynhyrchu nifer o ffurfiau anarferol ac annisgwyl (Thorne, 1996, tt. 255-275), (Williams S. J., 1980, tt. 119-113), gyda Jones (1988, t. 145) yn ei ddisgrifio fel '*a world of its own*'. Mae'n ferf afreolaidd sydd ddim yn dilyn patrwm rheolaidd berfau Cymraeg.

Mae'r ferf 'bod' yn newid yn sylweddol rhwng amseroedd a moddau gwahanol, sy'n aros yn gyson mewn ieithioedd eraill megis Saesneg (Williams S. J., 1980, tt. 119-120):

'bod'	**3ydd un.**
Mynegol Presennol	yw, ydyw, y mae, mae, oes
Mynegol Dyfodol a Phresennol Arferiadol	bydd
Mynegol Amherffaith	oedd
Mynegol Amherffaith Arferiadol	byddai
Mynegol Gorffennol	bu
Mynegol Gorberffaith	buasai
Dibynnol Presennol	bo, byddo
Dibynnol Amherffaith	bai, byddai
Gorchmynnol	bydded, boed, bid

Tabl 5 Rhediad y ferf afreolaidd, 'bod'

Fel mae Tabl 5 yn dangos, y ddau agwedd mwyaf anarferol yw ei ffurf trydydd person a'i allu i ddiflannu mewn rhai cymalau gofynnol (Borsley et al., 2007, tt. 256-261). Mae ffurf trydydd person presennol y ferf 'bod' yn ymddangos mewn cymalau 'os' (lle nad yw'n sicr) fel *yw/ydy* (pendant), neu *oes* (amhendant) yn unigol, ac *ydyn* yn y lluosog (Borsley et al., 2007, tt. 256-257), (Williams S. J., 1980, t. 129):

(7) a. *Nid **yw** Sioned yn aros.*
b. *A **yw** Sioned yn aros?*
c. *Os **yw** Sioned yn aros*

Yn ogystal â'r ffurfiau anarferol hyn, gall defnydd y ferf gael ei heffeithio gan amrywiadau yn iaith y siaradwr ei hunan, fel gyda threiglo uchod. Fel esbonnir isod yn 3, mae'r amrywiadau hyn yn ychwanegu at yr her prosesu sy'n barod yn gymhleth ar gyfer ieithoedd morffolegol gyfoethog. Mae'r ferf yn diflannu mewn rhai achosion ar lafar wrth iddi ymddangos ar ddechrau'r cymal er enghraifft, gan amlaf lle mae'r goddrych yn ail berson, unigol:

(8) a. *(Rwyt) ti'n mynd?*
b. *(Wyt) ti'n mynd, ynd wyt?*

Mae'n rhaid modelu'r cwestiwn cyfan a modelu'r cwestiwn heb ei ferf felly. Dengys Davies & Deuchar (2014) bod y ffurfiau gwahanol yn gallu cael eu heffeithio gan oedran y siaradwr a'i ddefnydd o Saesneg. Mae tuedd i ddileu'r ferf ar ddechrau cymalau fel hyn gan siaradwyr iau (o dan 50) sy'n defnyddio Saesneg yn aml ac felly am fodelu rhai o'u brawddegau mewn trefn debycach i Saesneg (SVO, mwy isod). Does dim patrwm pendant am hyn chwaith. Dengys astudiaeth Davies & Deuchar (2014, t. 226) bod rhai yn dileu'r ferf (57.66%) (Davies & Deuchar, 2010), a rhai ddim. Dydy bodolaeth un ddim felly yn dileu'r llall, ac mae rhai sefyllfaoedd lle mae'n rhaid i gyfrifiadur brosesu yn ôl cyd-destun. Mae angen ystyried geiriau sy'n cynhyrchu nifer o ffurfiau gwahanol fel hyn felly, a thalu sylw at amrywiadau ychwanegol posib hefyd wrth edrych ar iaith fwy sgyrsiol ac anffurfiol (mwy wrth drafod amrywiadau isod yn 3).

Trefn Geiriau o fewn Brawddeg

Gellid hefyd ystyried trefn geiriau o fewn brawddeg. Yn wahanol i weddill yr enghreifftiau yn y rhan hon, dydy trefn geiriau o fewn brawddeg ddim o reidrwydd yn cynhyrchu ffurfiau gwahanol, ond mae'n enghraifft o amrywio sy'n anarferol yn ieithoedd eraill y byd. Caiff ei gynnwys yn fras yn y rhan hon oherwydd ei fod yn dangos sut all nodwedd ramadegol fod yn her prosesu hefyd.

Yn gyffredinol, mae rhan helaeth o ieithoedd y byd wedi'u trefnu yn ôl trefn brawddeg *goddrych-gyntaf* (Dryer, 2013). Isod yn Ffigwr 14 gwelir adlewyrchiad rhyngwladol o drefn brawddeg ieithoedd y byd fesul canran (Tomlin, 1986, tt. 20-22).

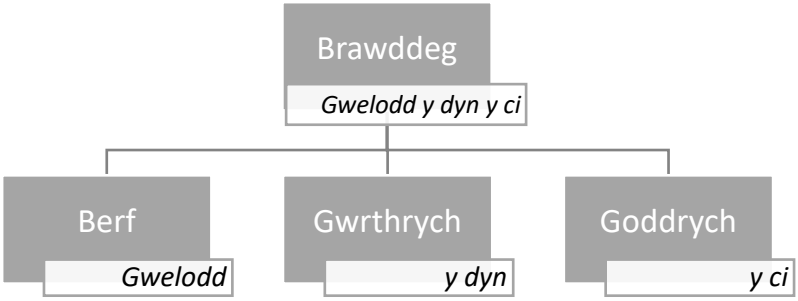
	SOV	SVO	VSO	VOS	OVS	OSV
% o ieithoedd y byd	44.78	41.79	9.2	2.99	1.24	0

Ffigwr 14 Trefn brawddeg ieithoedd y byd

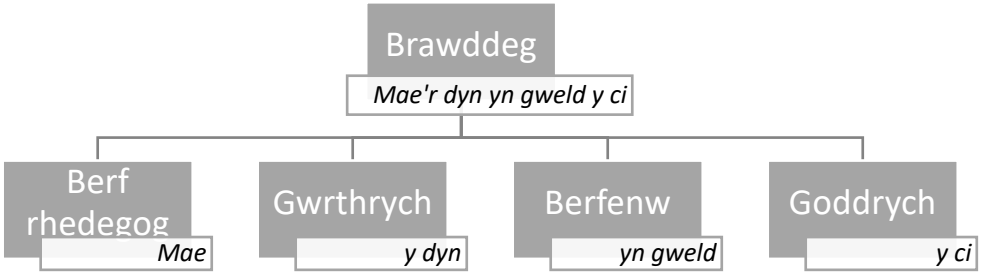
Mae'r Gymraeg yn anarferol yn y cyd-destun hwn gan ei fod yn gallu defnyddio trefn brawddeg *berf-oddrych-wrthrych* (*verb-subject-object,* VSO) sydd dim ond yn bodoli mewn 9.20% o ieithoedd y byd. Mae rhai ieithoedd eraill sy'n defnyddio'r drefn hon fel noda Davies & Deuchar (2010, tt. 82-83), lle mae modd creu brawddeg Saesneg sy'n *ferf-gyntaf* er enghraifft, '*Did the cat catch the mouse? Yes/No.'.* Yn y Gymraeg fodd bynnag, mae modd creu brawddeg sy'n defnyddio berfau wedi rhedeg fel yr uchod, neu gan ddefnyddio berfau cwmpasog, sy'n arbennig o gyffredin mewn iaith fwy sgyrsiol. Yma, mae ffurf rediadol y ferf, 'bod' gan fwyaf, yn dod o flaen y gwrthrych, ac mae'r berfenw wedyn yn ei ddilyn. Yn ogystal â hyn, mae Cymraeg yn dal i ddangos rhai olion o fod yn *ferf-ail* gyda strwythur y frawddeg annormal (isod, Ffigwr 15) fel noda Williams (1980, tt. 223-224) a Thorne (1996, t. 366).

Er nad yw trefn geiriau o fewn brawddeg yn cynhyrchu ffurfiau gwahanol mewn rhai ieithoedd, mae gwahanol ffyrdd o'i weld yn yr iaith Gymraeg. Fel yr enghreifftiau eraill yn y rhan hon, mae'n rhaid modelu ar gyfer pob un yn yr un ffordd. Gwelir yr enghreifftiau gwahanol o drefn geiriau o fewn brawddeg isod yn Ffigwr 15 (Awbery, 1976, t. 6).

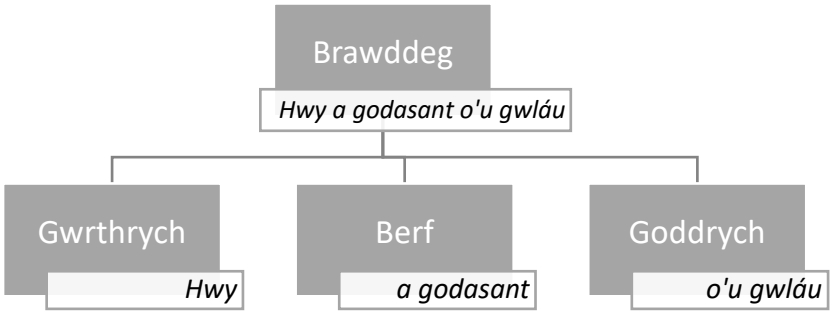
Brawddeg Berf Rediadol



Brawddeg Berf Gwmpasog



Brawddeg Berf Annormal



Ffigwr 15 Trefn brawddegau berf rediadol, gwmpasog, ac annormal

Ceir uchod gyflwyniad i rai enghreifftiau lle mae'r Gymraeg yn cynhyrchu nifer o ffurfiau gwahanol o un gair. Fel nodir uchod, mae'r Gymraeg yn iaith gynhyrchiol, sydd â nifer o eirffurfiau gwahanol yn deillio o'i morffoleg gyfoethog. Wrth ystyried y maes technoleg lleferydd, mae ieithoedd morffolegol gyfoethog dan anfantais ychwanegol lle bo diffyg data yn barod. I atgoffa, gyda mwy o ffurfiau gwahanol, mae angen mwy o ddata, a lle nad oes mwy o ddata ar gael, mae nifer uwch o eiriau OOV (allan o'r eirfa), sy'n arwain at fodelau yn methu adnabod gair a chanlyniadau gwaeth. Ffocysir ar yr

enghreifftiau uchod yn benodol oherwydd eu bod yn enghreifftiau anghyffredin neu'n wahanol i ieithoedd eraill y byd, a'u bod yn gallu amrywio eto rhwng siaradwyr yn dibynnu ar gyd-destun y sefyllfa. Ceir cyflwyniad i'r amrywiadau ychwanegol hyn isod, lle trafodir effaith amrywio cywair ac acen er enghraifft, ar ddefnydd siaradwyr Cymraeg o'r ffurfiau a geir uchod, a mwy. Arweinia hyn at ystyriaeth o leferydd 'gofalus' neu'n lleferydd 'naturiol' (Watkins A., 1961, t. 54), ac i ba fath o ddata dylid cynllunio a chasglu er mwyn cipio'r enghreifftiau hyn yn llwyddiannus wrth gyfathrebu â thechnoleg. Ceir ymchwiliad pellach i hyn yn y penodau i ddilyn.

3. Amrywiadau yn y Gymraeg

Pwrpas yr is-bennod hon yw ystyried amrywiadau gwahanol yn y Gymraeg all amrywio rhwng siaradwyr. Mae'r rhain yn enghreifftiau o amrywio sydd ddim ar sail morffoleg yr iaith fel uchod, ond yn hytrach yn dibynnu ar y siaradwr ei hun. Ceir yma rai enghreifftiau o amrywiadau sy'n amrywio rhwng lleferydd mwy gofalus a ffurfiol, a lleferydd cysylltiedig all fod yn fwy naturiol, ond yn anoddach i'w brosesu (rhesymau isod). Gwneir hyn er mwyn rhoi cipolwg ar y fath amrywiadau all godi mewn lleferydd mwy sgyrsiol, cyn ystyried amcanion y traethawd hir hwn yn 5, lle'r edrychir i fathau gwahanol o ddata all gipio'r amrywiadau hyn yn fwy llwyddiannus nag eraill.

Fel nodir yn Watkins (1961), mae modd defnyddio'r iaith Gymraeg yn 'ofalus' neu'n 'naturiol'. Y prif wahaniaeth rhwng y ddau yw defnydd o amrywiadau. Gall y rhain fod o ganlyniad i sefyllfa lle defnyddir cywair anffurfiol, lle mae rheolau gramadegol yn dueddol o lacio, a gellid gweld defnydd o gyfnewid cod lle cyflwynir geiriau Saesneg er enghraifft. Neu, gellid gweld amrywiadau gwahanol mewn sgwrs dafodieithol, lle defnyddir geiriau neu ynganiadau gwahanol, eto yn dibynnu ar y sefyllfa ac acen y siaradwr. Fel y geirffurfiau uchod, mae amrywiadau yn cyflwyno ffurfiau gwahanol o'r iaith ac felly'n ychwanegu at yr angen am fwy o ddata er mwyn cynnwys yr holl amrywiadau. Fel nodir uchod, lle nad oes digon o ddata i gynnwys y ffurfiau gwahanol, dydy'r model ddim yn gallu adnabod y gair ac mae'r canlyniadau'n gwaethygu. Adlewyrchir hyn yng ngwaith Wright (2006) lle cydnabyddir bod amrywiadau iaith rhwng siaradwyr yn her i brosesu ac mai'r ffordd amlycaf (er yn heriol i iaith lai ei hadnoddau) o ymdopi â'r her yw casglu mwy o ddata i adlewyrchu'r holl amrywiadau. Mae Wright hefyd yn cydnabod effaith ffactorau sosio-ieithyddol megis oedran, rhyw, a dosbarth cymdeithasol, all ddylanwadu ar ddefnydd siaradwyr o'r amrywiadau hyn. Ceir cyflwyniad i hyn ar gyfer y Gymraeg yn 3.1.

Isod, ffocysir ar amrywiadau sy'n deillio o amrywio cywair neu gyfnewid cod, ac acen a thafodiaith, gydag ystyriaeth o eilededdau sain. Gwneir hyn am eu bod yn rhai heriau prosesu cyffredin mewn Cymraeg cyfoes, all fod yn berthnasol i ieithoedd eraill llai eu hadnoddau hefyd.

3.1. Cywair

Caiff cywair ei ddiffinio mewn safon ISO (*International Organization for Standardization*) (2018) fel amrywiad sy'n cael ei ddefnyddio ar gyfer sefyllfaoedd gwahanol, yn enwedig ar gyfer graddfeydd gwahanol o ffurfioldeb. Mae'r safon ISO yn adnabod 18 cywair gwahanol, yn amrywio o sathredig neu isel (sefyllfaoedd bob dydd, nid ar bapur fel arfer), i anffurfiol (sefyllfaoedd cyfarwydd), i ffurfiol (sefyllfaoedd swyddogol). Mewn rhai cyd-destunau, gellir cyfeirio at diglossia hefyd. Ceir cyflwyniad i'r cysyniad hwn gan Thomas & Webb-Davies (2017, tt. 73-76), lle mae gan un math o iaith statws uwch, a bod disgwyliad i siaradwyr ddefnyddio un mewn sefyllfaoedd swyddogol, ac un arall yn deuluol er enghraifft. Arweinia hyn at siaradwyr yn siarad fersiynau gwahanol o'i hiaith mewn gwahanol sefyllfaoedd felly, a ddim yn siarad union yr un iaith bob tro. Ar gyfer y Gymraeg yn benodol, mae Prys (2016, t. 29) yn gwahaniaethu rhwng iaith frodorol (neu dafodieithol) ac iaith swyddogol Gymraeg fel rhai dewisiadau arddulliadol ar gyfer siaradwyr Cymraeg. Ar gyfer adnabod lleferydd, mae marcwyr ffonolegol a gramadegol yn gallu bod o gymorth wrth geisio mapio'r amrywiadau posib. Dengys y safon ISO (2018) bod geiriau mewn cywair anffurfiol yn aml yn cael eu hynganu'n llai clir ac yn gwyro i ffwrdd o reolau gramadegol (enghreifftiau isod). Fel nodir uchod, gall hyn fod yn her prosesu gan ei fod yn ychwanegu at yr angen am fwy o ddata er mwyn cynnwys yr holl amrywiadau posib; tasg sy'n barod yn heriol oherwydd anghenion data iaith morffoleg gyfoethog fel y Gymraeg, fel gwelir yn 2.

Un enghraifft o sut all yr iaith amrywio yn y modd hwn yw treiglo (2.1), lle mae siaradwyr yn gallu bod yn llai tebygol o ddefnyddio ffurfiau treigledig mewn cywair anffurfiol (ISO, 2018, tt. 16-17). Mae gwaith Prys (2016, t. 42) yn atgyfnerthu hyn, yn nodi bod llai o ddefnydd o dreiglo wrth ddefnyddio cywair anffurfiol, ond nad oes patrwm clir am ba ddefnydd sy'n goroesi yn y fath sefyllfaoedd. Mae enghreifftiau pellach yn Prys (2016, t. 39), gydag amrywiadau wrth negyddu: 'nid wyf yn mynd', 'dw i (dd)im yn mynd', wrth ddefnyddio'r geiryn perthynol 'a': 'beth a glywaist?', 'be glywaist di?', ac eraill. Enghraifft arall yn y Gymraeg yw amrywiad mewn rhagenwau lle

defnyddir y clitig *fy* + rhagenw ôl *i* yn ffurfiol, ond ddim yn anffurfiol (Borsley et al., 2007, tt. 158-159): 'fy nghar i ydy hwnna', 'car fi 'dy hwnna', ac wrth ddefnyddio arddodiaid rhediadol (Borsley et al., 2007, t. 308), fel gwelir isod yn Tabl 6. Dylid nodi fodd bynnag nad yw'r gwahaniaeth hwn yn ddu a gwyn hollol, fel noda Davies (2016); mae amrywiaeth eang yn yr iaith anffurfiol lle mae defnydd o'r ffurf 'ffurfiol' yn gallu ymddangos wrth siarad yn anffurfiol hefyd.

	gan	
	Ffurfiol	**Anffurfiol**
1 un.	gennyf i	gynna i
2 un.	gennyt ti	gynna ti
3un. (g)/(b)	ganddo ef/ganddi hi	gynno fo/gynni hi
1 ll.	gennym ni	gynno ni
2 ll.	gennych chi	gynno chi
3 ll.	ganddynt nhw	gynnyn nhw

Tabl 6 Amrywiad cywair anffurfiol mewn arddodiad rhediadol

Yn ogystal â'r ffurfiau gwahanol sydd i'w gweld yn 2 felly, mae angen ystyried modelu ar gyfer yr amrywiadau pellach yma hefyd, os am gyfrif am bob ffurf posib yr iaith Gymraeg wrth ei phrosesu.

Sosio-ieithyddiaeth ac Amrywiadau Arddulliadol

Fel nodir uchod, mae gwaith Wright (2006) yn cydnabod bod amrywiadau iaith rhwng siaradwyr yn her i brosesu, ond bod hefyd angen ystyried ffactorau sosio-ieithyddol sydd yn ymwneud â chefndir y siaradwyr. Mae gwaith ymchwil sosio-ieithyddiaeth yn dangos bod, er enghraifft, oedran, rhyw, a dosbarth cymdeithasol, yn gallu cael effaith ar y ffordd mae pobl yn siarad, sydd felly hefyd yn cael effaith ar adnabod lleferydd. Dengys astudiaeth Misnikov (2019) ar rai o'r systemau adnabod lleferydd sydd ar gael ar y farchnad, Apple Siri a Google Home, eu bod yn ymateb i rai amrywiadau ieithyddol rhwng siaradwyr Saesneg Americanaidd, ond bod cyfyngiadau i'w heffeithiolrwydd wrth ddelio ag amrywiadau arddulliadol, hyd yn oed yn Saesneg. Rhan o beth sy'n achosi arferion gwahanol rhwng siaradwyr yw eu cefndir sosio-ieithyddol wrth siarad. Dengys gwaith Labov ar Martha's Vineyard (1963) ac Efrog Newydd (2006) sut all sefyllfa sosio-ieithyddol siaradwr, er enghraifft eu hoedran, rhyw a'u dosbarth cymdeithasol, effeithio

ar eu defnydd o amrywiadau mewn iaith. Gan amlaf, mae astudiaethau yn y maes wedi darganfod yr isaf y dosbarth cymdeithasol, neu'r lleiaf ffurfiol yw'r sefyllfa, y mwyaf yw defnydd arddull mwy amrywiol. Yng nghyd-destun y Gymraeg, damcaniaeth Coupland & Ball (1989, t. 11) yw bod hyn oherwydd nad oedd yr iaith yn cael ei ddefnyddio o fewn addysg na darlledu tan yn lled ddiweddar.

Wrth gasglu data sain ar gyfer y project SpeechDat Cymru, mae Jones et al. (1998, t. 3) yn nodi bod oedran yn arbennig yn gallu effeithio ar natur llais y siaradwr Cymraeg a'u hynganiad a chystrawen. Mae gwaith Davies (2016, t. 32) yn esbonio bod amrywiad iaith yn ôl oedran yn gallu bod yn arwydd o newid iaith, yn adlewyrchu sut mae iaith wedi newid dros amser. Mae gwaith Davies (2016) a Davies & Deuchar (2014) yn astudio'r ffenomenon hyn wrth astudio nodweddion gramadegol penodol (dilead rhan o'r ferf gynorthwyol a chystrawennau meddiannol), sy'n dangos mwy o amrywiadau 'anhraddodiadol' rhwng siaradwyr iau (2016, t. 32). Mae Holmes (2001, tt. 150-174) yn tynnu sylw at rôl ystyriaeth arall, sef rhyw, o fewn amrywiadau ieithyddol, wrth nodi bod menywod yn aml yn siarad yn wahanol i ddynion er nad ydynt yn ymwybodol o wneud bob tro. Mae Jones et al. (1998, t. 3) yn tynnu sylw at sut mae llais menyw yn dueddol o fod â thraw uwch a nodweddion mwy anadlol (*breathy*) i'w lleisiau er enghraifft. Agwedd arall bwysig i'w ystyried yw dosbarth cymdeithasol, fel nodir uchod. Mae hyn i wneud â bri (*prestige*) wrth geisio ymddangos yn fwy parchus, neu'n rhan o ddosbarth cymdeithasol uwch (Prys, 2016, tt. 27-28). Adlewyrchir hyn yng ngwaith Labov (2006) hefyd, fel nodir uchod, lle edrychodd ar amrywiadau ieithyddol gweithwyr mewn siopau o ardaloedd gwahanol yn Efrog Newydd. Yn gyffredinol, wrth fynd i ardaloedd lle'r oedd tebygrwydd uwch o fod ymhlith pobl o ddosbarth cymdeithasol uwch, roedd llai o amrywiadau ieithyddol yn ymddangos. Wrth ystyried adnabod lleferydd yn benodol, gwelir crynodeb Clark et al. (2019) bod gweithiau yn y gorffennol wedi edrych ar y ffordd y mae pobl yn siarad â chyfrifiaduron a sut mae hwn yn amrywio o sgwrs gyda pherson arall. Cyfeirir yn benodol at waith Amalberti et al. (1993) sy'n dangos, o brofion lle cymharir grwpiau gwahanol o bobl wrth siarad â chyfrifiadur neu â pherson, bod pobl yn defnyddio iaith fwy gofalus wrth siarad â chyfrifiadur. Gwelir hefyd waith Koulouri et al. (2016) sy'n edrych ar aliniad ac yn darganfod bod mewnbwn ac allbwn tebycach yn arwain at ganlyniadau gwell, sy'n tynnu ar yr hyn a drafodir ym mhennod II 4 Dulliau i Ymdopi gyda Llai o Ddata.

Yng nghyd-destun adnabod lleferydd, mae tuedd siaradwyr o amrywio'u defnydd o'r ffurfiau uchod yn 2, yn ôl cywair y sefyllfa, yn ychwanegu at yr her o brosesu iaith yn gyffredinol.

3.2. Cyfnewid Cod

Mae safon ISO (ISO, 2018, t. 14) yn diffinio cyfnewid cod fel ffenomenon gyffredin mewn cymunedau dwyieithog lle mae geiriau ac ymadroddion o un iaith yn cael eu cynnwys mewn iaith arall, gan fwyaf ar lafar. Maent yn nodi fod hyn yn gysylltiedig â chywair anffurfiol, ymysg siaradwyr iau, yn tynnu ar yr uchod. Mae Ritchie & Bhatia (2012) yn adleisio hyn, bod siaradwyr yn cyfnewid ieithoedd yn aml mewn sefyllfaoedd gwahanol, yn dibynnu ar y siaradwyr eraill, y pwnc, a lleoliad y sgwrs (2012, t. 378). Caiff ei gynnwys yn y rhan hon oherwydd ei fod yn un enghraifft bellach o sut mae angen casglu mwy o ddata eto er mwyn cynnwys amrywiadau rhwng siaradwyr, ar ben y ffurfiau gwahanol uchod yn 2. Gellir eu gweld yn aml wrth ddefnyddio geiriau sathredig i regi, yn aml er mwyn lleihau neu ddwysáu effaith y rhegi. Dylid nodi bod modd defnyddio cyfnewid cod mewn sefyllfaoedd ffurfiol mewn cymunedau eraill, er enghraifft ymysg cymunedau yn Ne Affrica, lle mae cyfnewid cod i Saesneg yn gyffredin mewn trafodaethau academaidd (ISO, 2018, t. 14). Fodd bynnag, mae agwedd siaradwyr Cymraeg tuag at gyfnewid cod i Saesneg yn gallu bod yn negyddol fel noda Prys (2016, t. 30), gan bwysleisio dylanwad Saesneg ar y Gymraeg yn hanesyddol, yn ogystal ag ymgyrchoedd i'w leihau yn y cyfryngau dros y blynyddoedd. Mae peth gwaith, megis Amazouz et al. (2017) gyda Ffrangeg-Arabeg, wedi darganfod bod modd trin y modelau a rhannu corpora mewn ffyrdd penodol er mwyn ymdopi ac amrywiadau cywair mewn adnabod lleferydd, ond mae'n gofyn am ddatblygu a phrosesu ychwanegol. Gweler hefyd waith Lyu & Lyu (2008) ar Mandarin-Taiwanese lle defnyddir tri set gwahanol o ddata o gyfuniadau o'r ddwy iaith, gwaith Lyu et al. (2015) ar Mandarin-Saesneg lle casglwyd corpws 93 awr o leferydd cymysg o'r ddwy iaith, a gwaith Modipa et al. (2013) ar Sepedi-Saesneg lle defnyddiwyd dau eiriadur ynganu i greu un system adnabod lleferydd dwyieithog. Gyda'r holl enghreifftiau hyn, gwelir bod angen gwaith datblygu ychwanegol neu/a maint mwy o ddata er mwyn ymdopi gyda'r amrywiadau ychwanegol.

Wrth gasglu data felly, gall fod mantais wrth ystyried sut mae'r Gymraeg yn newid o dan ddylanwad yr iaith Saesneg. Mae Cymru yn swyddogol yn wlad ddwyieithog ac mae gan Saesneg statws swyddogol ochr yn ochr; mae'r Gymraeg yn newid felly, yn rhannol oherwydd dylanwad Saesneg (Prys, 2016, t. 30). Mae'r plethu yma rhwng

Saesneg a'r Gymraeg wedi arwain at gyfnewid cod. Mae Thomas & Webb-Davies (2017, t. 47) yn trafod sut gall siaradwyr dwyieithog (er enghraifft, Cymraeg a Saesneg), gyfuno ieithoedd wrth sgwrsio. Mae system adnabod lleferydd felly, angen gallu adnabod hyn hefyd os yw siaradwyr am ddefnyddio cyfnewid cod wrth gyfathrebu drwy dechnoleg lleferydd. Gellir rhannu cyfnewid cod fewn i iaith matrics ac iaith wedi'i mewnblannu (Deuchar & Davies, 2009, t. 20), lle mae'r iaith matrics yn dynodi'r morffemau system a'r iaith wedi'i mewnblannu yn dynodi'r morffemau cynnwys, er enghraifft, 'y *thing* yma'. Yma, gwelir strwythur Cymraeg gydag enw Saesneg yn y cynnwys. Felly, yr iaith matrics yw Cymraeg lle'r iaith wedi'i mewnblannu yw Saesneg. Yn ôl Deuchar & Davies (2009, t. 34), mae hwn yn achos clasurol o gyfnewid cod, sy'n awgrymu sefydlogrwydd dwyieithog. Gan fwyaf, mae cyfnewid cod yn ymddangos fel tuedd anffurfiol yn y Gymraeg (Deuchar & Davies, 2009, t. 19), (Thomas & Webb-Davies, 2017, t. 60), sy'n ymddangos mewn sefyllfaoedd lle siaredir mewn cywair anffurfiol. Er hyn, fel nodir uchod, mae cyfnewid cod yn digwydd mewn sefyllfaoedd ffurfiol hefyd fel gwelir yng ngwaith Prys (2016). Gwelir o ganlyniadau Deuchar et al. (2016) bod siaradwyr brodorol sydd wedi dysgu Saesneg a Chymraeg ochr yn ochr fel babanod, yn cyfnewid cod fwyaf aml.

Mewn cymuned ddiglossig ddwyieithog Cymraeg a Saesneg felly, ceir amrywiadau ar ben y ffurfiau gwahanol a geir yn 2. Fel nodir uchod, mae gwaith ar gymunedau dwyieithog eraill wedi gweld canlyniadau addawol wrth drin y modelau a rhannu'r corpora mewn ffyrdd penodol, i greu math o system ddwyieithog neu aml arddulliadol. Fodd bynnag, mewn cyd-destunau llai eu hadnoddau, mae gofyn am fwy o waith datblygu neu am fwy o ddata yn heriol.

3.3. Tafodiaith ac Acen

Enghraifft arall bwysig i'w chynnwys yn y rhan hon yw ystyriaeth o dafodiaith ac acen. Ceir ymchwiliad ehangach i effaith yr agweddau hyn yn benodol ym mhennod VII Amrywio Iaith. Caiff ei chynnwys yn y rhan hon gan ei bod yn agwedd sy'n gyffredin iawn yn y Gymraeg ond hefyd mewn ieithoedd eraill llai, a mwy, eu hadnoddau. Mae gwaith Awbery (1996) yn cyflwyno tafodiaith fel math o amrywiad ieithyddol sy'n newid o'r naill ardal i'r llall, ac yn effeithio ar y ffordd y mae pobl yn siarad a'r geiriau maent yn eu defnyddio. O fewn tafodiaith, mae Awbery yn awgrymu bod amrywiad acen yn nodwedd ddylanwadol hefyd wrth edrych ar amrywio rhwng siaradwyr, yn cwmpasu

amrywio ynganiad a geirfa'r siaradwyr, eto yn ychwanegu ffurfiau gwahanol sydd angen mwy o ddata i'w prosesu'n llwyddiannus.

Ar gyfer adnabod lleferydd, mae amrywiadau seiniau mewn acenion yn faes sydd wedi ennyn diddordeb gan dimoedd ymchwil, yn enwedig ar gyfer amrywiadau Saesneg. Gweler er enghraifft waith cynnar Humphries & Woodland (1997), a gwaith Tjalve & Huckvale (2005), ill dau yn edrych ar ddulliau yn defnyddio nodweddion acenion penodol er mwyn gwella effeithiolrwydd systemau Saesneg America o adnabod Saesneg Prydain. Gweler gwaith Najafian et al. (2014), sy'n edrych ar effaith acenion tafodieithol Saesneg Prydain ar berfformiad system adnabod lleferydd, ac yn darganfod bod graddfeydd gwallau ar gyfer lleferydd sydd ag acenion trwm, gogleddol fel Swydd Efrog er enghraifft, yn gallu bod saith gwaith yn fwy na lleferydd heb amrywiadau o'r fath. Ceir enghraifft o waith ar acen De-Affrica Saesneg yng ngwaith Kamper et al. (2012), lle defnyddir dull modelu ar gyfer acenion penodol er mwyn gweld gwelliant addawol. Mewn trosolwg o'r sefyllfa, gyda ffocws ar Fandarin, noda Huang et al. (2004) bod eu harbrofion ar acenion gwahanol o fewn systemau adnabod lleferydd yn dangos bod acen yn her sylweddol ar gyfer y maes, gan eu bod yn anodd i'w mapio'n glir a bod diffyg data. Un o brif broblemau'r maes wrth brosesu acenion yw bod graddfa acen person mor eang, a bod acenion yn newid yn ôl lle mae person yn cael eu magu, yn symud i fyw, â phwy maent yn siarad yn aml, ac yn y blaen. Gwelir isod rhai enghreifftiau ar gyfer tafodieithoedd Cymraeg yn benodol, o sut gall acenion gwahanol amrywio iaith siaradwr.

<u>Amrywiadau geirfa</u>

Heb drafod amrywiadau geirfa, fe awgrymir, yn anghywir, bod tafodiaith ac acen yn effeithio seiniau yn unig. Mae geirfa yn gallu amrywio fesul acen a siaradwr, fel dengys gwaith Willis (2019) ar ddefnydd y gair 'cau' er enghraifft, sydd felly yn gallu ychwanegu at yr her prosesu cyfrifiadurol, fel nodir uchod. Mae rhai amrywiadau geirfaol yn dilyn patrwm de a gogledd ond â rhai eithriadau, fel gwelir yn Thomas & Thomas (1989, tt. 13-15), er enghraifft y parau isod.

(9) a. mas allan
b. tad-cu taid
c. mam-gu nain
ch. nawr rŵan
d. allwedd goriad

Er bod modd rhannu'r uchod mewn i de ar y brig a gogledd ar y gwaelod, does dim ffin bendant ar fap gwaetha'r modd. Mae Rees (2016, t. 48) wedi gwneud ymchwil i geisio darganfod lle mae'r newid yn digwydd: tua gogledd Ceredigion lle defnyddir y ddau. Ceir rhagor o waith ar y pwnc gan Thomas & Thomas (1989, tt. 13-16) gan edrych ar y gair am 'melysion' yn Gymraeg: 'los(h)in(s)' yn y de, a 'fferi(n)s' yn y gogledd – y 'gogledd' i'w weld uwchben yr afon Rheidiol a'r 'de' i hen siroedd Penfro, Caerfyrddin, a Morgannwg yn y cyd-destun hwn. Dylid nodi bod modd cael mwy nag un gair â'r un ystyr hefyd er enghraifft 'budur' a 'brwnt' i feddwl angen golchi (Thomas & Thomas, 1989, t. 21). Wrth gwrs, mae modd i 'brwnt' feddwl 'garw' (*rough*) mewn rhai tafodieithoedd gogleddol hefyd, yn debyg i ystyron gwahanol y gair 'clwyd' fel 'esgynbren' (yr ysgol y bydd ieir yn clwydo arni) neu 'llidiart' (cyfystyr â 'clwyd') fel noda Thomas & Thomas (1989, t. 19), sy'n her arall wrth brosesu lleferydd mewn tafodieithoedd gwahanol. Noda Thomas & Thomas (1989, tt. 21-22) drydydd amrywiad geirfaol sef geiriau arbenigol, sy'n digwydd o fewn ardal arbennig er enghraifft, 'wtra' am 'ffordd yn arwain i fferm'. Dylid cydnabod bod gwaith Thomas & Thomas uchod yn deillio o'r 1960au ac felly mae'r nodweddion ieithyddol dan sylw yn cynrychioli iaith o ddiwedd yr 19eg ganrif mewn sawl achos. Yn wir, mae'r ffaith fod geiriau mewn iaith yn newid dros amser yn her ychwanegol eto i dechnoleg iaith fel adnabod lleferydd.

Yn ddelfrydol, byddai modd casglu enghreifftiau o'r holl ffurfiau hyn er mwyn eu modelu'n gyfrifiadurol ond, mae hon yn dasg anferth ac felly yn arbennig o heriol mewn cyd-destun llai ei hadnoddau fel y Gymraeg.

<u>Amrywiadau seiniau</u>

Yng nghyd-destun modelau acwstig ar gyfer system adnabod lleferydd, mae amrywiadau seiniau yn gallu bod yn allweddol i fodelu. Dyma felly, yw ffocws yr ymchwil ehangach ym mhennod VII Amrywio Iaith. Mae'r amrywiadau yn gallu bod yn her prosesu oherwydd cymhlethdodau ychwanegu fersiynau gwahanol o'r un gair i eiriadur ynganu. Un o'r prif wahaniaethau rhwng seiniau yn acenion y de a'r gogledd yng Nghymru er enghraifft, yw amrywiad yn eu hynganiad o rai llafariaid. Ceir cyflwyniad i lafariaid y Gymraeg yng ngwaith Mayr & Davies (2011) fel /i: ɪ e: ɛ a a: o: ɔ u: ʊ ə/ gyda dau gategori o Gymraeg gogleddol fel /ɨ:/ ac /ɨ/. Maent hefyd yn cefnogi gweithiau Awbery (1984), Jones (1984) a Ball & Williams (2001) wrth nodi bod Cymraeg gogleddol yn gwahaniaethu rhwng 13 o gategorïau deuseiniaid (*diphthongs*) gwahanol, lle dim ond 8

sydd ar gyfer yr acen ddeheuol. Fel nodir uchod, mae'r amrywiadau gwahanol hyn yn her prosesu. Wrth gasglu data sain ar gyfer adnabod lleferydd Cymraeg, atgyfnerthodd Jones et al. (1998, t. 4) sut mae hyd llafariaid yn gallu peri problem wrth gynllunio data sain ar gyfer y pwrpas hwn.

Caiff y ffomem /ɨ/ ei gategoreiddio gan Awbery (2009, tt. 359-360) a Thomas & Thomas (1989, t. 31), fel ffonem ychwanegol yn y gogledd, fel yr 'u-ogleddol'. Gwelir y gwahaniaeth yn y parau isod (Rees, 2016, t. 48):

(10) a. 'mul' /mɨːl/ 'mil' /miːl/
b. 'sur' /sɨːr/ 'sir /siːr/
c. 'tŷ' /tɨː/ 'ti' /tiː/

Yn gyffredinol, i siaradwr gogleddol, byddai'r pâr yn cael eu hynganu'n wahanol (gwelir trawsgrifiadau uchod). Er hyn, byddai siaradwr ag acen ddeheuol yn ynganu'r parau gyda'r ynganiad /iː/ yn unig, a byddai'n rhaid dibynnu ar gyd-destun er mwyn deall y frawddeg. Dylid nodi bod eithriadau i'r diffiniad 'gogleddol'; gwelir rhestr gyflawn yn Thomas & Thomas (1989, t. 31), yn ogystal â chyfnewid rhwng y ddau mewn rhai ardaloedd, fel gwelir yng ngwaith Rees (2016), (2015). Gwelir amrywiadau pellach ynghylch ansawdd llafariaid sy'n amrywio yn ôl y seiniau cyfagos yn Ball (1984), Thomas & Thomas (1989, tt. 34-35), ac Awbery (1984):

(11) a. /coːsb/ /cɔsb/, 'cosb'
b. /coːst/ /cɔst/, 'cost'
c. /coːɬ/ /cɔɬ/, 'coll'
ch. /paːsg/ /pasg/, 'pasg'
d. /peːɬ/ /pɛɬ/, 'pell'
dd. /gweːɬt/ /gwɛɬt/, 'gwellt'
f. /gweːɬ/ /gwɛɬ/, 'gwell'

Yn fras, ceir yr ynganiad deheuol ar y chwith, a'r gogledd ar y dde, ond mae rhai eithriadau i hyn lle mae'r sain hirach yn fwy cyffredin yn y gogledd (1984, tt. 71-72) (enghreifftiau isod).

(12) a. /cʊsg/ /cuːsg/, ‘cwsg’
b. /sʊllt/ /suːɬt/, ‘swllt’

Wrth i fwy o ddysgwyr fynd ati i ddysgu’r iaith Gymraeg fel ail iaith, mae seiniau newydd yn cael eu trosglwyddo i’r iaith (Rees & Morris, 2018, t. 41), ac felly mae angen ystyried hyn hefyd wrth gynllunio data adnabod lleferydd. Gyda mwy o bobl yn dysgu’r iaith, mae hwn yn gallu arwain at ynganiadau ac amrywiadau newydd, fel gwelir yng ngwaith Ellis (2008, tt. 349-350), Miesel et al. (1981) ac Ortega (2009, t. 51). Dros amser mae’r amrywiadau hyn yn gallu gwanhau, ac maent yn mabwysiadu ynganiadau mwy brodorol (rhyngiaith, (Selinker, 1972)) neu, ar adegau, mae’r amrywiadau’n aros yn iaith y siaradwr ail iaith am byth (ffosileiddio, (Han, 2004, t. 4)). Wrth amrywio ynganiadau ymhellach, cyflwynir ffurfiau gwahanol eto, sy’n gallu arwain at fodelau yn methu adnabod geiriau yn iawn.

Eilededdau Sain

Diffinnir eilededdau sain (*alternations*) gan Ball & Williams (2001, tt. 50, 54-55) fel achos lle mae sain neu gyfres o seiniau yn cael ei ynganu’n wahanol mewn lleferydd cysylltiedig a gofalus. Gall yr ynganiad amrywio i fod yn debycach i sain arall, ac yng nghyd-destun lleferydd, mae hwn yn cyfeirio at effaith seiniau cyfochrog. Mae cymathu (*assimilation*) yn gysylltiedig yma hefyd, wrth gyfeirio at seiniau sydd ddim yn cael eu hynganu o gwbl mewn lleferydd cysylltiedig, gan amlaf o fewn clwstwr cytseiniaid lle collir un aelod o’r clwstwr. Nodir uchod mai’r prif wahaniaeth rhwng y ddau fath o leferydd yw defnydd o amrywiadau o fewn lleferydd cysylltiedig, ac fe ystyrir effaith amrywiadau seiniau ar fathau gwahanol o ddata adnabod lleferydd fel eilededdau sain ymhellach yn y penodau i ddilyn. Wrth amrywio ynganiad y sain neu’r seiniau, crëir ffurfiau gwahanol unwaith eto, sy’n ychwanegu at yr angen am fwy o ddata er mwyn eu cynnwys. Fel nodir uchod, gall hyn fod yn heriol ar gyfer iaith lai ei hadnoddau lle mae diffyg data yn gyffredin yn barod.

Mae eilededdau yn digwydd gan amlaf gyda seiniau gorfannol (*alveolar*) neu ddeintiol (*dental*) er enghraifft /t, d, n/, lle maent yn newid i fod yn ddwywefusol (*bilabial*) neu felar (*velar*). Gwelir yr enghreifftiau isod yn Ball & Williams (2001, tt. 51-52).

(13) a. /ˈkrid mawr/ > /ˈkrib ˈmaʊr/, ‘crud mawr’
b. /mɛʊn ˈkudɪn/ > /mɛʊŋ ˈkudɪn/, ‘mewn cwdyn’

Dylid nodi bod modd i eilededdau ddigwydd lle mae clwstwr o ddwy sain orfannol ar derfyn gair, neu lle mae lleisio mewn geiriau cyfansawdd, yn aml mewn geiriau sy’n cychwyn gyda ‘Llan-’, ac eraill. Gwelir yn Ball & Williams (2001, tt. 53-54) bod modd ei weld yn digwydd ar draws ffiniau geiriau hefyd. Mae hyn yn arbennig o berthnasol wrth ystyried mathau gwahanol o ddata, fel a geir ym mhennod VI Math o Ddata. Yn fwy penodol eto, ceir enghreifftiau o seingolli hefyd, sy’n debyg ond yn colli’r sain yn hollol yn hytrach na’i ymdebygu i sain arall, er enghraifft yr isod (2001, tt. 54-55).

(14) a. /ˈanadl/ > /ˈanal/, ‘anadl’
b. /fɛnɛstr/ > /fɛnɛst/, ‘ffenestr’

Caiff hyn ei weld yn aml hefyd mewn geiriau sy’n gorffen gyda /v/ neu /ð/, neu wrth golli’r sain /r/. Ceir hefyd agwedd arall ar eilededdau, sef sillgolli, sy’n cyfeirio at golled sillafau cyfan yn hytrach na dim ond seiniau, er enghraifft yr isod.

(15) a. /əɪˈsdɛðvəd/ > /ˈsdɛðvəd/, ‘eisteddfod’
b. /naˈdɔlɪg/ > /dɔlɪg/, ‘Nadolig’

Gwelir rhain yn aml ymysg acenion gogleddol yn ôl Awbery (2001, t. 55), ond mae modd eu gweld yn acen Cwm Tawe hefyd, ac eraill. Mae’r agwedd hon yn arbennig o berthnasol wrth ystyried effaith amrywiadau gwahanol o ganlyniad i acen, a gaiff ei drafod ymhellach ym mhennod VII Amrywio Iaith.

Ar gyfer prosesu cyfrifiadurol, mae eilededdau sain yn gallu bod yn her, lle mae angen ystyried ffurfiau gwahanol y gair wrth gael eu hynganu fel rhan o linyn o eiriau wedi’u cysylltu (neu leferydd cysylltiedig), yn hytrach na geiriau unigol. Wrth fynd ati i gasglu data geiriau unigol fel a geir yng nghorpws Paldaruo ym mhennod III 2.3 GALLU, Paldaruo a Macsen, mae’n bosib bod ffurfiau sy’n amrywio mewn lleferydd cysylltiedig yn cael eu colli wrth i siaradwyr ynganu un gair ar y tro. Caiff hwn ei brofi ym mhennod VI Math o Ddata sy’n ystyried mathau gwahanol o ddata i ymchwilio i hyn. Fel gyda nifer o’r amrywiadau uchod, ac fel mae Ball & Williams (2001) yn nodi, mae’r iaith yn amrywio yn gyson a does dim consensws clir i ba raddau mae’r amrywiadau hyn yn

systematig yn ôl yr iaith ei hun, neu os ydynt yn amrywiadau arddulliadol gan y siaradwr. Yn ddelfrydol felly, dylid ymchwilio ymhellach i effaith lleferydd cysylltiedig ar amrywiadau fel eilededdau sain, all gael eu colli wrth gasglu data ar ffurf geiriau unigol yn unig. Dylid ceisio cynnwys holl ffurfiau posib yr iaith wrth gasglu data hyfforddi adnabod lleferydd, boed hynny yn ffurfiau gwahanol ar sail morffoleg gyfoethog yr iaith ei hun, neu'n amrywiadau gan siaradwyr unigol. Fel nodir uchod, gall hyn fod yn her prosesu sylweddol wrth ystyried sefyllfa ieithoedd llai eu hadnoddau lle mae data yn aml yn brin. Yn y penodau i ddilyn, ceir ystyriaeth o'r mathau gwahanol o ddata sydd orau i ymdopi gyda'r heriau prosesu hyn, lle bo diffyg data yn her yn barod.

4. Casgliadau

Mae'r bennod hon yn cynnig cyflwyniad i eirffurfiau ac amrywiadau gwahanol y Gymraeg, ac amlinellu sut mae ffactorau ychwanegol sosio-ieithyddol yn gallu amrywio lleferydd siaradwr ymhellach. Prif nod y bennod hon yw dangos sut mae ffurfiau gwahanol yn gallu bod yn her prosesu gan eu bod yn ychwanegu at yr angen am fwy o ddata er mwyn eu cynnwys yn y data hyfforddi. Lle bo data yn brin, fel sydd yn wir ar gyfer ieithoedd llai eu hadnoddau fel y Gymraeg, gall hyn ychwanegu at y nifer o eiriau sy'n ymddangos yn yr OOV (allan o'r eirfa), ac arwain felly at fodelau gwannach a chanlyniadau gwaeth.

Gwelir yn 1.1 bod datblygiadau yn y maes wedi'u dylunio'n helaeth ar gyfer ieithoedd mwy eu hadnoddau fel Saesneg, a bod hyn wedi amharu ar y gefnogaeth sydd ar gael ar gyfer ieithoedd llai eu hadnoddau fel y Gymraeg. Gall hyn arwain at heriau ychwanegol wrth i ieithoedd llai eu hadnoddau, sy'n aml yn forffolegol gyfoethog, fynd ati i gasglu data ar gyfer adnabod lleferydd. Ceir cyflwyniad i rai geirffurfiau gwahanol yn 2, sy'n enghreifftiau o sut mae'r Gymraeg yn iaith gynhyrchiol sy'n creu nifer o ffurfiau gwahanol ar sail ei morffoleg. Ceir yn 3 enghreifftiau pellach o amrywiadau iaith sy'n gallu creu ffurfiau gwahanol yn ôl iaith y siaradwr ei hun. Yn y rhan hon, mae enghreifftiau o dafodiaith ac acen yn enwedig o berthnasol ar gyfer ystyriaeth pennod VII Amrywio Iaith, lle ffocysir ar effaith amrywiadau acen ar system adnabod lleferydd Cymraeg. Mae'r effaith o leferydd cysylltiedig yn cael ei brofi ym mhennod VI Math o Ddata.

Fel nodir uchod, y prif reswm bod hyn oll yn her prosesu yw oherwydd ei fod yn ychwanegu at yr angen am ddata neu fodelau mwy, sy'n barod yn heriol ar gyfer

ieithoedd llai eu hadnoddau. Ceir ym mhennod VI Math o Ddata, ystyriaeth o fathau gwahanol o ddata gellid mynd ati i'w casglu a'u defnyddio, sy'n gwneud yn fawr o ddata llai, er mwyn ceisio ymdopi gyda'r ffurfiau gwahanol hyn. Gwelir yr amcanion a chwestiynau ymchwil sy'n arwain ymlaen o hyn oll yn y rhan nesaf, 5. Mae'r penodau nesaf yn amlinellu'r profion a gynhelir i ymchwilio'r cwestiynau hyn.

5. Amcanion yr Ymchwil

Mae tri phrif amcan i'r ymchwil yn y traethawd hir hwn er mwyn deall a phrofi heriau ieithyddol sy'n wynebu ymchwilwyr yn y maes. Yn gyntaf, er mwyn gosod sylfaen i brofion pellach, edrychir i ba ffordd o fodelu'r Gymraeg yn gyfrifiadurol sy'n creu'r system adnabod lleferydd mwyaf cadarn. Gwneir hyn drwy gymharu canlyniadau'r system wrth brosesu'r un set o ddata, corpws Paldaruo, yn gyntaf o fewn cit offer sy'n defnyddio modelau ystadegol, HTK, o fewn cit offer croesryw sy'n defnyddio cyfuniad o fodelau ystadegol a niwral, *Kaldi*, ac yna o fewn peiriant sy'n defnyddio modelau pen-i-ben, *DeepSpeech*. Yr ail amcan yw edrych i ba faint a math o ddata sydd fwyaf effeithiol o fewn system adnabod lleferydd Cymraeg. Gwneir hyn drwy gymharu'r system wrth brosesu tri fersiwn gwahanol o gorpws Paldaruo, yn cynrychioli tri gwahanol maint o'r un data. Crëir promptiau newydd hefyd, ar ffurf newydd brawddegau llawn, er mwyn cymharu gyda'r ffurf geiriau unigol sydd i'w weld yng nghorpws Paldaruo. Defnyddir platfform *Common Voice-cy* er mwyn casglu'r data newydd ar ffurf brawddegau llawn ac yna, cymharu'r canlyniadau hynny gyda'r rheiny a gyfrifwyd o'r profion blaenorol ar gorpws Paldaruo yn unig. Yn drydydd, mae'r traethawd hir hwn hefyd yn edrych i effaith amrywio yn yr iaith ar adnabod lleferydd Cymraeg, ac yn edrych i weld sut mae amrywiadau acen yn benodol yn effeithio system adnabod lleferydd. Defnyddir data wedi'i gyfrannu gan S4C ar gyfer creu setiau profi acen benodol ar gyfer y de a'r gogledd. Profir y data hwn ar system adnabod lleferydd wedi'i hyfforddi gan ddata Paldaruo a data *Common Voice-cy*, yn ogystal â data S4C ei hun, er mwyn deall mwy am effaith acen ar adnabod lleferydd. Gan fod ffactorau eraill megis cywair, cyfnewid cod, neu dafodiaith, yn aml yn effeithio ar acen siaradwr, edrychir ar effaith y ffactorau hyn hefyd. Yn olaf, arbrofir gyda gwahanol setiau data a ffurfweddau o'r peiriant *DeepSpeech* er mwyn gweld a oes modd gwella'r effeithiolrwydd o gwbl wrth brosesu'r Gymraeg. Gwneir hyn gan fod angen llai o arbenigedd ieithyddol er mwyn ei ddefnyddio, a bod y maes rhwydweithiau niwral (sydd yn cael eu defnyddio gan fodelau pen-i-ben

DeepSpeech) yn newydd i ymchwil ieithoedd llai eu hadnoddau, a bod ymchwil ar y pwnc ar gyfer y Gymraeg yn brin.

Fel nodir ym mhennod II 1.2 Defnyddiau o Dechnoleg Lleferydd, mae ymchwil yn dangos bod adnabod lleferydd yn gallu bod yn fanteisiol ym meysydd archifo, addysg, ac iechyd. Gwelir hyn yng ngwaith Lanchantin et al. (2013) a Mohr et al. (2013) ar archifau'r BBC, a gwaith Jong et al. (2018) ar archifau yn yr Iseldiroedd, yn ogystal â gwaith Carrier (2017) ar gyfer addysg plant ysgol, a Neri et al. (2003), (2003) i ddysgwyr ail iaith. Mae modd gweld defnydd y dechnoleg hon ar gyfer cleifion sy'n dioddef o glefydau megis Parkinson a Dementia yng ngwaith Delič et al. (2014), Mulvenna et al. (2017), a Jilbab et al. (2019). Yr anfantais yn hyn i gyd yw bod rhan helaeth y datblygiadau yma wedi'u dylunio ar gyfer ieithoedd mwy fel Saesneg. Fel nodir uchod, mae ymchwil yn y maes technoleg lleferydd yn brin ar gyfer ieithoedd llai eu hadnoddau fel y Gymraeg.

Amlinellir uchod rai o'r manteision o gael technoleg o'r fath ar gyfer ieithoedd eraill, ac felly gwelir yr angen ar gyfer adlewyrchu'r datblygiadau ar gyfer y Gymraeg. Gwelir potensial y gwaith hwn ym mhroject Lleisiwr yr Uned Technolegau Iaith (2018), sy'n bancio lleisiau cleifion yn synthetig, cyn iddynt eu colli o gyflyrau megis Clefyd Niwronau Echddygol a Chanser y Gwddf er enghraifft. Roedd hwn yn adeiladu ar fuddsoddiadau mewn ymchwil yn y maes testun-i-leferydd (TIL) gan sefydliadau megis Llywodraeth Cymru. Fel nodir yn I Cyflwyniad, mae eu strategaethau a'u cynlluniau gweithredu mwyaf diweddar megis *Cymraeg 2050: Miliwn o siaradwyr* (2017) a *Chynllun gweithredu technoleg Cymraeg* (2018), yn amlinellu'r manteision sydd i'w gweld wrth gefnogi technoleg lleferydd y Gymraeg. Maent yn pwysleisio'r angen am fuddsoddiad mewn seilwaith technoleg iaith ar raddfa fawr (2017, tt. 70-71) ac ar ddatblygu sail dystiolaeth gadarn drwy werthuso ac ymchwil (2017, tt. 77-78). Nod y gefnogaeth hon yw sicrhau bod modd i siaradwyr Cymraeg ddefnyddio'r iaith mewn cynifer o sefyllfaoedd â phosib, er mwyn sicrhau dyfodol yr iaith yn wyneb datblygiadau pwysig yn y byd technolegol.

Nod y traethawd hir hwn felly yw adeiladu ar y wybodaeth ynghylch adnabod lleferydd yn benodol, a'r heriau o ran data ac amrywio yn yr iaith sydd yn gallu effeithio ar lwyddiant adnabod lleferydd. Mae ffocws ar fodelau gwahanol, maint a math y data, ac amrywio yn yr iaith, er enghraifft acen, gan fod y ffactorau hyn yn effeithio ar ieithoedd llai eu hadnoddau eraill tu allan i Gymru, yn ogystal â'r Gymraeg. Mae'n gyfle i gymryd trosolwg o'r sefyllfa fel y mae hi, ac ymchwilio sut i'w gwella. Gan fod y Gymraeg yn

iaith leiafrifol, a bod technoleg iaith yn dechnoleg allweddol ar gyfer galluogi (*enabling technology*) (Evas, 2013, t. 5), mae'n hollbwysig darparu argymhellion i fod yn ganllaw i strategaethau, ymchwil, a masnach y dyfodol. Mae'r traethawd hir hwn felly yn ceisio darparu trosolwg o'r maes fel mae'n sefyll a chynnig awgrymiadau am sut i ddatblygu a gwella. Ceir isod y cwestiynau ymchwil.

Wrth ymdopi gydag iaith lai ei hadnoddau sy'n dioddef o ddiffyg data fel y Gymraeg,

1. Pa ffordd o fodelu'r Gymraeg yn gyfrifiadurol sy'n creu'r system adnabod lleferydd mwyaf cadarn?
2. Pa faint a math o ddata sydd fwyaf effeithiol o fewn system adnabod lleferydd Cymraeg?
3. Sut mae amrywio acen yn effeithio system adnabod lleferydd Cymraeg?

Wrth ystyried ymchwil sydd wedi'i gyflawni ar adnabod lleferydd ar gyfer ieithoedd llai eu hadnoddau eraill, disgwylir i fodelau croesryw greu system fwy cadarn (canlyniadau gwell), na modelau ystadegol wrth brosesu'r Gymraeg. Mae hyn oherwydd bod modelau croesryw, fel a geir yng nghit offer *Kaldi*, yn gwneud yn fawr o ddulliau diweddar rhwydweithiau niwral, heb golli manteision dulliau ystadegol megis defnydd gwybodaeth ieithyddol, a gaiff eu defnyddio gan git offer ystadegol HTK. Disgwylir na fydd modelau pen-i-ben, fel *DeepSpeech*, yn creu system fwy cadarn na'r hyn a grëir gan fodelau croesryw, gan fod modelau pen-i-ben yn dibynnu'n fwy ar faint y data gan nad oes defnydd gwybodaeth ieithyddol. Ceir cyflwyniad i'r citiau offer yn y bennod nesaf, V Citiau Offer. Gan fod diffyg data ar gyfer y Gymraeg, disgwylir i hyn fod yn her sylweddol i lwyddiant y modelau pen-i-ben.

O ran maint a math y data, disgwylir na fydd hyfforddi a phrofi ar faint mwy o'r un data dro ar ôl tro yn effeithiol yn yr hir dymor, gan na fydd y system yn gwella'r gallu i adnabod lleferydd newydd, dim ond ei allu i adnabod set gyfyngedig o ddata. Wrth edrych ar fath y data, mae'n debyg y bydd data ar ffurf brawddegau llawn yn gwella effeithiolrwydd y system wrth gyfrifo tebygolrwydd y lleferydd mae'n adnabod, gan ei fod yn rhoi mwy o gyd-destun ffonolegol (eilededdau sain ac amrywiadau eraill) i'r geiriau a'r seiniau yn y mewnbwn na data ar ffurf geiriau unigol.

O edrych ar amrywio acen yn y Gymraeg, disgwylir y bydd y ffurfiau gwahanol a grëir gan acenion gwahanol, a'r ffactorau ychwanegol megis cywair, cyfnewid cod, a

thafodiaith, yn cael effaith mawr ar ganlyniadau system adnabod lleferydd wrth ymdopi gyda diffyg data oherwydd her prosesu ffurfiau gwahanol y Gymraeg. Yn fwy penodol, gan fod mwy o gyfranwyr data Paldaruo 4.0 o'r gogledd nac o'r de, ac felly bod mwy o ddata gogleddol yn y corpws, bydd y set brofi ogleddol yn debygol o roi canlyniadau gwell na'r set brofi ddeheuol. Yn gyffredinol fodd bynnag, lle mae cydbwysedd acenion ymysg cyfranwyr, disgwylir i'r acen ogleddol gael effaith negyddol ar y canlyniadau oherwydd bod seiniau ychwanegol ynddi o'i chymharu ag acen y de, er enghraifft yr 'u-ogleddol' fel nodir uchod. Yn ychwanegol i hyn, mae'n debyg bydd defnyddio setiau hyfforddi a phrofi tebyg, o'r un math o ddata neu o'r un corpws penodol, yn cael effaith gadarnhaol ar y canlyniadau, oherwydd bod y system yn cael ei phrofi ar ddata tebyg i'r data hyfforddi.

Dyma'r astudiaeth gyntaf o effaith y ffactorau hyn ar adnabod lleferydd yn y Gymraeg ac felly, bydd casgliadau'r traethawd hir hwn yn arwain at ddealltwriaeth gwell o'r maes. Ar gyfer ymchwilwyr yn y maes, bydd yn cynnig canllawiau am sut i fynd â'r ymchwil ymhellach a sut i ystyried ymdopi gydag amrywio ieithyddol. Ar gyfer sefydliadau cyllido, bydd yn cynnig mewnwelediad i'w buddsoddiadau ar waith, a sut orau i barhau yn y dyfodol; ar gyfer y gymuned ieithoedd llai eu hadnoddau, bydd yn cynnig argymhellion am sut i fynd â'u hymchwil hwythau ar adnabod lleferydd ymhellach, gan dynnu ar wersi a ddysgwyd wrth arbrofi gyda'r Gymraeg.

Rhennir y cwestiynau ymchwil hyn i bedwar set o brofion, i ddilyn yn y bedair pennod nesaf. Ceir ym mhennod V Citiau Offer fanylion profion i gymharu modelau ystadegol a chroesryw, yn ogystal ag ystyriaeth o feintiau gwahanol o ddata. Mae pennod VI Math o Ddata yn cymharu mathau gwahanol o ddata: geiriau unigol a brawddegau llawn, cyn mynd ati i ystyried dylanwad amrywio acen ym mhennod VII Amrywio Iaith. Yn olaf, mae pennod VIII *DeepSpeech* yn ystyriaeth o'r dulliau modelu mwyaf diweddar yn y maes. Mae'r penodau i ddilyn yn rhoi manylion y profion perthnasol ac yn cyflwyno'r canlyniadau ynghyd â thrafodaeth fer. Ceir trafodaeth gyffredinol ym mhennod IX Trafodaeth.

V. Citiau Offer

1. Cyflwyniad

Pwrpas y bennod hon yw amlinellu'r profion a'r data a ddefnyddiwyd er mwyn ymchwilio cwestiwn ymchwil 1 (IV 5 Amcanion yr Ymchwil), ynghylch cymharu gwahanol ffyrdd o fodelu'r Gymraeg wrth ymdopi gyda diffyg data (gwelir pennod I 2.4 Diffyg Data, Amrywio Ieithyddol a'r Dulliau Perthnasol). Mae hefyd yn edrych i gwestiwn ymchwil 2 ynghylch pa faint o ddata sydd orau ar gyfer y defnydd hwn. Pwrpas y profion hyn yn benodol yw deall beth yw effaith defnyddio modelau ystadegol a modelau croesryw, er mwyn darganfod pa ffordd sydd orau o fodelu'r Gymraeg, a beth yw effaith ychwanegu mwy o'r un data, ar gyfer adnabod lleferydd. Yn y bennod hon, ceir esboniad o'r profion a gynhelir er mwyn profi ymarferoldeb y rhagdybiaethau perthnasol sydd i'w gweld yn 1.4. Mae'r bennod wedi'i rhannu i dair rhan yn dilyn y cyflwyniad: Dull (Data, Gweithdrefn), Canlyniadau a Thrafodaeth. Bydd trafodaeth gyffredinol am oblygiadau ac arwyddocâd y canlyniadau ym mhennod IX Trafodaeth.

1.1. Yr Her: Diffyg Data

Mae'r profion yn y bennod hon yn ffocysu ar fodelu wrth ystyried diffyg data. Fel nodir ym mhennod I 2.4 Diffyg Data, Amrywio Ieithyddol a'r Dulliau Perthnasol, mae diffyg data yn gyffredin ymysg ieithoedd llai eu hadnoddau, fel y Gymraeg. Mae'n cyfeirio at ieithoedd lle nad oes digon o adnoddau addas ar gael, lle mae 'digon' yn cyfeirio at ddata mawr (tua 10,000 awr, ffigwr yn amcangyfrif o faint corpora cwmnïoedd masnachol Apple/Amazon). Mae'r term 'iaith lai ei hadnoddau' yn cyfeirio at iaith sydd heb fynediad at ddata digonol, a all fod yn broses hir, drud, a chymhleth sy'n gofyn am arbenigwyr a phŵer cyfrifiadurol. Mae hefyd yn cyfeirio at ieithoedd lle nad oes llawer o ddeunydd ar gael dan drwydded agored (gwelir pennod III 1.2 Trwyddedu). Gwelir cymhariaeth rhwng adnoddau agored Cymraeg a Saesneg isod yn Tabl 7:

	Corpws	Oriau
Saesneg	TED-LIUM	452
	Librispeech	tua 1,000
Cymraeg	Paldaruo 4.0	38
	Common Voice-cy (13/02/19)	42

Tabl 7 Cymharu nifer oriau corpora agored Saesneg a Chymraeg

Dyma rhai o'r corpora agored Saesneg (TED-LIUM (Hernandez et al., 2019) a *Librispeech* (Panayotov et al., 2015)) y defnyddiodd cwmni Mozilla (III 2.4 Mozilla, *Common Voice* a *Common Voice* Cymraeg) wrth greu eu system adnabod lleferydd cyntaf (Ardila, et al., 2020; Morais, 2017). Roedd cyllid ychwanegol ar gael i Mozilla er mwyn cael mynediad at filoedd o oriau pellach drwy Fisher (tua 2,000 awr) (Cieri et al., 2004) a *Switchboard* (260 awr) (Godfrey et al., 1992) hefyd. Mae ymchwil wedi dangos yn glir bod mwy o ddata yn gwella canlyniadau adnabod lleferydd (Lamel, Gauvain, & Adda, 2002; Lamel, Gauvain, & Adda, 2002; Moore, 2003; Moore, 2001), gyda Moore yn awgrymu byddai angen rhwng 3,000,000 a 10,000,000 awr o ddata er mwyn cyrraedd adnabod lleferydd hollol lwyddiannus gyda 0% WER (cyfradd geiriau gwallus) (Moore, 2003, t. 2584). Wrth greu system Saesneg, mae mwy o argaeledd arbenigedd, cyllid, a data nac wrth greu system ar gyfer iaith fel y Gymraeg. Dylid nodi nad yw ffigurau Tabl 7 yn cynrychioli'r meintiau mwy sy'n cael eu defnyddio gan gewri masnachol y maes, Amazon neu Apple, sydd siŵr o fod tua neu'n uwch na 10,000 awr (amcangyfrif oherwydd rheolau diogelwch a phreifatrwydd y cwmnïau).

Mae'r profion yn y bennod hon wedi'u dylunio i gymharu ymarferoldeb modelau ystadegol fel a geir yng nghit offer HTK, a modelau croesryw fel a geir yng nghit offer *Kaldi*. Hynny yw, gyda diffyg data yn realiti, pa fodel sy'n rhoi canlyniadau gwell dan yr amgylchiadau.

1.2. Cymharu HTK a *Kaldi*

Prif rôl cit offer yw hyfforddi modelau er mwyn prosesu lleferydd. Ceir cyflwyniad i adeiladwaith modelau yn gyffredinol ym mhennod II 3 Modelau, lle mae esboniad o'r gwahaniaethau rhwng model ystadegol a model pen-i-ben yn ogystal â chroesryw sy'n gymysgedd o'r ddau (manylion isod yn 2.2). I atgoffa, mae modelau ystadegol HMM-GMM a chroesryw HMM-DNN yn ddulliau o adeiladu model wrth ddefnyddio gwybodaeth ieithyddol megis geiriadur ynganu (manylion isod yn 2.2). Mae'r wybodaeth ieithyddol hwn yn rhoi trawsgrifiad o'r geiriau mewn lecsicon i helpu gyda'r modelu, ond yn gallu bod yn broses ddrud ac araf gan fod angen arbenigwyr ieithyddol i'w ddatblygu'n effeithiol (manylion project arbenigol ym mhennod III 2.3 GALLU, Paldaruo a Macsen). Mae'r bennod hon yn cymharu citiau offer sy'n defnyddio modelau ystadegol a chroesryw. Mae HTK yn git offer sy'n defnyddio modelau ystadegol HMM-GMM. Mae *Kaldi* yn defnyddio cymysgedd o fodelau ystadegol HMM-GMM a chroesryw HMM-DNN. Dylid pwysleisio, er bod y ddau git offer yn defnyddio technegau gwahanol er mwyn adeiladu'r modelau, mae'r ddau'n dal i ddefnyddio gwybodaeth ieithyddol. Gwelir cymhariaeth gyda pheiriant sy'n gwneud i ffwrdd â'r wybodaeth ieithyddol ym mhennod VIII *DeepSpeech*.

Fel nodwyd ym mhennod III 2.3 GALLU, Paldaruo a Macsen, gwelir llwyddiant wrth ddechrau datblygu adnabod lleferydd Cymraeg gydag HTK yn 2015. Ond, wrth i *Kaldi* amlygu fel opsiwn mwy diweddar, oedd yn boblogaidd yn y gymuned dechnolegol ac yn cynnig manteision dros HTK (er enghraifft, gallu creu modelau acwstig *ac* iaith; HTK dim ond yn creu modelau acwstig), defnyddiwyd hwn yn ei le yn natblygiad nesaf y project yn 2016-17 (III 2.3 GALLU, Paldaruo a Macsen). Fel nodir ym mhennod II 3 Modelau, gwelir poblogrwydd *Kaldi* yn ei ddefnydd rhyngwladol megis ar gyfer yr ieithoedd Hindi (Upadhyaya et al., 2017), (Sri et al., 2020) a Punjabi (Guglani & Mishra, 2018). Mae ymchwil wedi'i gynnal hefyd sy'n cymharu citiau offer gwahanol yn erbyn ei gilydd wrth brosesu lleferydd o ieithoedd mwy. O'r citiau offer sy'n god agored, mae *Kaldi* yn cael ei gymeradwyo'n gyson wrth iddo gynhyrchu canlyniadau addawol wrth brosesu lleferydd Saesneg (Matarneh et al., 2017), ac Almaeneg (Gaida, et al., 2014) er enghraifft. Cynhelir y profion yn y bennod hon er mwyn cymharu canlyniadau citiau offer HTK a *Kaldi* wrth brosesu lleferydd Cymraeg, sy'n iaith lai ei hadnoddau yn dioddef o ddiffyg data.

1.3. Meini Prawf

Er mwyn deall mwy cyn dechrau cymharu canlyniadau'r profion, defnyddir y meini prawf isod (Tabl 8) er mwyn cymharu symlrwydd, trwydded ac ymarferoldeb HTK a *Kaldi* ar gyfer y Gymraeg.

	HTK	***Kaldi***
Symlrwydd	• Modiwlau adeiladu; • **Modiwlau hyfforddi**; • Cod C++.	• Modiwlau hyfforddi; • Cod C++ ond â **ryseitiau/sgriptiau parod**.
Trwydded	• Trwydded i Microsoft ac Ysgol Peirianneg Prifysgol Caergrawnt (HTK 2006); • Cod lled agored gyda rhai cyfyngiadau e.e. methu ailddosbarthu na dileu unrhyw hawlfraint; • Hawl i'w ddefnyddio ond ddim i'w ailddosbarthu; • Hawl datblygu academaidd ond nid masnachol.	• Trwydded i Apache (Apache, 2019); • **Cod hollol agored** sy'n hybu arloesedd ac yn cynnig cymorth gan gymuned o ddatblygwyr a defnyddwyr; • Hawl i'w ddefnyddio, ei ailddosbarthu; • Hawl datblygu academaidd a masnachol.
Ymarferoldeb	• Angen lot o ddata; • Angen arbenigedd datblygu ac ieithyddol.	• Angen lot o ddata; • Angen arbenigedd datblygu ac ieithyddol; • **Defnyddio dulliau mwy diweddar**; • **Cefnogaeth cymuned dechnolegol eang *Kaldi*.**

Tabl 8 Cymharu HTK a *Kaldi* o'r meini prawf

O'r meini prawf hyn yn Tabl 8, *Kaldi* sy'n ymddangos fel y cit offer fwyaf addas ar gyfer y dasg hon. Gwelir bod HTK yn cynnig mwy o fodiwlau hyfforddi na *Kaldi* ond, wrth gydbwyso'r anfantais hon gyda mantais ryseitiau a sgriptiau parod sydd ar gael yn *Kaldi*, ymddengys mai *Kaldi* yw'r cit offer mwyaf syml i'w ddefnyddio. Mae *Kaldi* yn cynnig cod hollol agored lle nad yw HTK yn gwneud. Mae'r agwedd hon yn hollbwysig ar gyfer

ieithoedd llai eu hadnoddau, lle mae pwysau i greu effaith (*impact*) yn ogystal â diffyg adnoddau cyllid yn aml. Er mwyn gwneud hyn yn effeithiol, mae angen y drwydded fwyaf agored a chaniataol a phosib (Prys & Jones, 2015). Gan fod angen lot o ddata ac arbenigedd ar y ddau git offer, mae mantais dulliau mwy diweddar (manylion isod yn 2.2), a chefnogaeth y gymuned dechnolegol sydd y tu ôl i *Kaldi* yn ei arwain at y brig yn y gymhariaeth hon (Tabl 8).

Mae'r profion yn y bennod hon yn edrych yn agosach ar effeithiolrwydd modelau ystadegol, fel a geir yn HTK, a modelau croesryw fel a geir yn *Kaldi*, wrth adnabod lleferydd Cymraeg. Mae'r profion yn ffocysu ar HTK a *Kaldi* oherwydd mai dyma'r citiau offer sydd wedi'u defnyddio cyn belled ar gyfer y Gymraeg, am eu bod yn rhad ac am ddim, ac yn cyfrif fel trwydded agored (HTK yn fwy cyfyngedig na *Kaldi*, III 1.2 Trwyddedu). Mae'r meini prawf uchod yn rhoi'r argraff mai *Kaldi* sydd orau oherwydd ei fod wedi adeiladu ar y seilwaith greodd HTK, a'i ehangu wrth ychwanegu datblygiadau newydd yn y maes i mewn i'r system (II 3.2 Modelau Croesryw).

1.4. Cwestiynau Ymchwil a Rhagdybiaethau Perthnasol

I atgoffa, y cwestiynau ymchwil dan sylw yn y bennod hon yw rhif 1 a 2. Mae'r cyntaf ynghylch pa ffordd o fodelu'r Gymraeg yn gyfrifiadurol sy'n creu'r system adnabod lleferydd mwyaf cadarn. Fy rhagdybiaeth gyntaf yw y bydd adeiladwaith croesryw cit offer *Kaldi* yn rhoi canlyniadau gwell nac HTK wrth ymdopi gyda diffyg data, gan fod adeiladwaith croesryw *Kaldi* yn gwneud yn fawr o ddatblygiadau diweddar rhwydweithiau niwral, lle nad yw HTK yn gwneud.

Defnyddir fersiynau data corpws Paldaruo (Cooper et al., 2018), (Cooper et al., 2019) er mwyn cynnal y profion. Ynghylch yr ail gwestiwn ymchwil, defnyddir y profion hyn i ystyried os yw ychwanegu maint mwy o'r un data yn ddull effeithiol o wella'r canlyniadau. Fy ail ragdybiaeth yn y bennod hon yw bod ychwanegu data yn gwella'r canlyniadau, ond na fydd hyfforddi a phrofi ar faint mwy o'r un data dro ar ôl tro yn effeithiol yn yr hir dymor. Fy namcaniaeth am hyn yw y byddai'r system yn adnabod y set gyfyngedig hynny o ddata yn unig.

2. Dull

Bydd y rhan yma yn disgrifio'r dull o hyfforddi a phrofi modelau yn y ddau git offer: HTK a *Kaldi*. Yn 2.1, ceir cyflwyniad i ddata corpws Paldaruo a'r fersiynau gwahanol o'r corpws a ddefnyddir yn y profion i gymharu canlyniadau'r citiau offer wrth ymdopi gyda data Cymraeg. Yn 2.2, esbonnir adeiladwaith y citiau offer. Er mwyn mynd i'r afael â chwestiwn ymchwil 1, ynghylch modelau cyfrifiadurol (IV 5 Amcanion yr Ymchwil), ceir cyflwyniad i'r dull k-*fold cross-validation* (C-V) yma, sydd wedi profi'n fanteisiol gyda data llai wrth leihau tuedd (Jurafsky & Martin, 2019, t. 69), (Cawley & Talbot, 2010, t. 2102). Ceir hefyd esboniad o'r WER (cyfradd geiriau gwallus) a W.Acc (cyfradd geiriau cywir) sy'n cael eu defnyddio i fesur effeithiolrwydd y citiau offer.

2.1. Data

Defnyddir data o gorpws Paldaruo (Cooper et al., 2018), (Cooper et al., 2019) er mwyn profi pa git offer, a pa fath o fodel felly, sy'n well wrth ymdopi gyda data llai. Mae'r data yng nghorpws Paldaruo yn ddata sain Cymraeg sydd wedi'i gynllunio'n arbennig ar gyfer prosesu lleferydd. Mae'r data yn seinegol gytbwys sy'n golygu ei fod yn adlewyrchu defnydd mwyaf cyffredin o seiniau'r iaith Gymraeg (Prys & Jones, 2018). Cynlluniwyd y data yn 2014 fel rhan o'r project GALLU yn Uned Technolegau Iaith, Canolfan Bedwyr. Ceir esboniad o gefndir y project hwn ym mhennod III 2.3 GALLU, Paldaruo a Macsen. Casglwyd y data hwn drwy ddefnyddio'r dull torfoli, a lwyddodd i gasglu recordiadau gan wahanol gyfranwyr. Gwelir metadata y cyfranwyr isod yn Tabl 9 o fersiwn 4.0 y corpws. Mae peth metadata gan gyfranwyr ar goll yn Tabl 9, fel gwelir wrth gymharu gyda Tabl 10. Mae gan y rhan fwyaf o'r cyfranwyr acenion gogleddol (41.62%). Roedd rhan fwyaf y cyfranwyr dan 40 oed (61.60%) ac roedd canran uchel yn categoreiddio'u hunain fel acen iaith gyntaf yn hytrach nac acen dysgwr/ail iaith (86.54%).

Rhedodd y project hwn am 6 mis ac fe ddatblygwyd promptiau geiriau unigol seinegol gytbwys, i gipio'r seiniau mwyaf cyffredin yn yr iaith. Cafodd ei gyflawni drwy ddylunio promptiau recordio ar gyfer cyfranwyr i ddarllen yn uchel (Cooper et al., 2019, t. 9). Ar ôl dylunio promptiau i gynnwys holl seiniau, neu gyfuniadau o seiniau, mwyaf cyffredin yr iaith, aethpwyd ati i drefnu'r promptiau er mwyn hwyluso'r dasg o ddarllen yn uchel. Gwelir y ddwy enghraifft ganlynol o brompt geiriau unigol corpws Paldaruo hefyd ym mhennod III 2.3 GALLU, Paldaruo a Macsen.

(16) a. LLEUAD, MELYN, AELODAU, SIARAD, FFORDD, YMLAEN, CEFNOGAETH, HELEN

b. RHYBUDDIO, ELEN, UWCHRADDIO, HWNNW, BEIC, CYMRU, RHOI, AELOD

Mae'r traethawd hir hwn yn manteisio ar yr agwedd hon o gorpws Paldaruo ym mhennod VI Math o Ddata sy'n edrych ar ba fath o ddata sy'n fwyaf addas ar gyfer adnabod lleferydd Cymraeg: geiriau unigol neu frawddegau llawn.

Oed	**Nifer**	**%**	**Rhyw**	**Nifer**	**%**
18-30	163	31.35%	Benyw	260	50.00%
31-40	157	30.25%	Gwryw	260	50.00%
41-50	88	16.96%			
51-60	65	12.52%			
61-70	35	6.74%			
71-80	10	1.93%			
80+	1	0.19%			
Cyfanswm	519	99.94%		520	100.00%

Lleoliad Acen	**Nifer**	**%**	**Math Acen**	**Nifer**	**%**
Canolbarth	75	14.42%	Dysgwr/ail iaith	70	13.46%
De-ddwyrain	75	14.45%	Iaith Gyntaf	450	86.54%
De-orllewin	101	19.46%			
Gogledd-ddwyrain	52	9.83%			
Gogledd-orllewin	217	41.62%			
Cyfanswm	520	99.78%			100.00%

Tabl 9 Metadata cyfranwyr corpws Paldaruo (fersiwn 4.0)

Fel gwelir yn Tabl 10 isod, 38 awr o ddata sain sydd yn y corpws ar ei fwyaf, sy'n cyfrif fel data llai. Cyfeirir at drothwy data 'digonol' ym mhennod I 2.4 Diffyg Data, Amrywio Ieithyddol a'r Dulliau Perthnasol. Cyhoeddwyd y corpws hwn am y tro cyntaf ar 30/11/14 fel fersiwn 1.0, oedd yn cynnwys 26 awr gan 383 cyfrannwr. Ers hynny, mae fersiynau 2.0, 3.0, a 4.0 wedi eu cyhoeddi. Gan fod corpws Paldaruo wedi parhau i gasglu data, cyhoeddwyd fersiwn 5.0 ar 19/12/18, oedd ar ôl cyfnod profi'r profion hyn; roedd hwnnw'n cynnwys 40 awr gan 564 cyfrannwr. Roedd fersiwn 5.0 hefyd yn cynnwys

promptiau newydd ar ffurf brawddegau llawn felly ni fyddai'n brawf teg o'i gymharu gyda fersiynau cynt (mwy o fanylion ym mhennod VI Math o ddata). Gwelir manylion y fersiynau gwahanol isod yn Tabl 10.

Fersiwn Paldaruo	2.0	3.0	4.0
Cyhoeddwyd	15/07/2016	09/06/17	06/11/17
Amser Sain (awr)	34	36	38
Nifer o Ffeiliau	11,556	12,024	12,682
Nifer o Gyfranwyr	487	506	536

Tabl 10 Fersiynau corpws Paldaruo

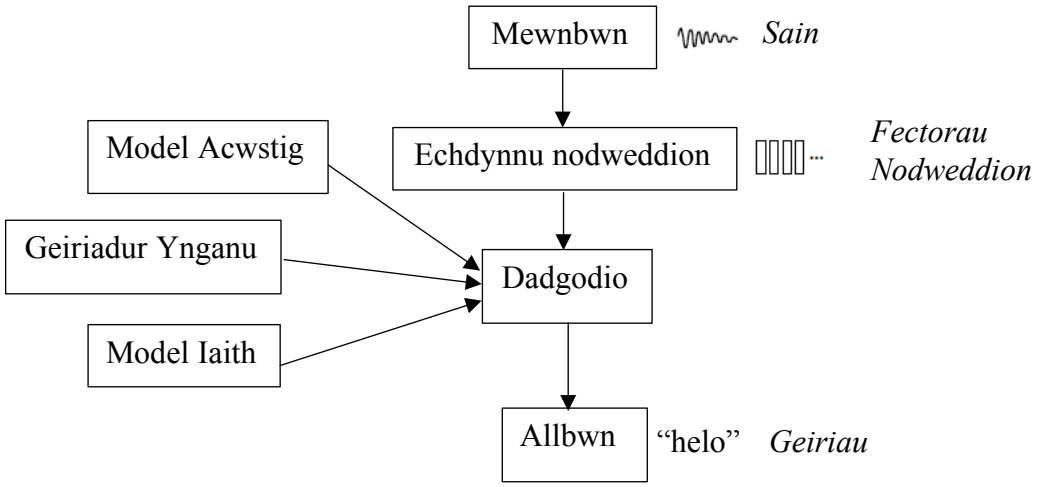
(wedi'i addasu o (Cooper, Jones, & Prys, 2019, t. 6))

Ychwanegwyd promptiau pellach ar ffurf cwestiynau yn 2015 er mwyn cefnogi project arall o fewn yr Uned Technolegau Iaith (Cooper et al., 2019, t. 5). Caiff y rhain eu hidlo o'r data a ddefnyddir yn y profion hyn er mwyn sicrhau prawf teg wrth gymharu mathau o ddata ym mhennod VI Math o Ddata sy'n cymharu geiriau unigol a brawddegau llawn.

Wrth brofi'r data hwn o fewn y citiau offer HTK a *Kaldi*, mae modd cymharu eu heffeithiolrwydd. Bydd canlyniad y profion hyn yn edrych i fy rhagdybiaeth uchod (1.4) mai modelau croesryw *Kaldi* sydd fwyaf addas ar gyfer defnydd o fewn adnabod lleferydd Cymraeg.

2.2. Gweithdrefn

Prif rôl cit offer mewn system adnabod lleferydd yw hyfforddi modelau sy'n cyfrannu at y broses adnabod lleferydd, fel gwelir yn Ffigwr 16 isod. Mae'r profion hyn yn ffocysu ar effaith gwahanol newidolion ar fodelau acwstig. Caiff y model iaith ei hyfforddi gyda'r testun o'r corpws gaiff ei ddefnyddio ar bob cam. Y citiau offer sy'n hyfforddi'r modelau acwstig wrth echdynnu nodweddion seinegol o'r signal sain er mwyn dadgodio'r signal hwnnw (Jurafsky & Martin, 2019, tt. 534-539), a hyfforddi'r system i ragdybio beth sy'n cael ei ddweud ar sail y signal sain hwnnw (Ffigwr 16), (Jones et al., 2019, tt. 26-27), (Jurafsky & Martin, 2019, tt. 548-563).



Ffigwr 16 Pensaernïaeth cit offer yn fras

(wedi'i addasu o (Gales & Young, 2007, t. 201))

Ceir esboniad o'r broses hyfforddi a phrofi yn fras yn y rhestr isod, hynny yw o'r mewnbwn sain i'r allbwn geiriau.

Paratoi Data

1. addasu sgriptiau i fod yn addas ar gyfer y dull C-V;
2. llwytho fersiwn o gorpws Paldaruo i lawr o wefan Porth Technolegau iaith Cenedlaethol Cymru;
3. Creu ffeil .csv (*comma separated values*) ar gyfer bob sampl;
4. golygu, hidlo, chwynnu o fewn Excel (data o brosiectau amherthnasol e.e. promptiau Wicipedia, Macsen, Lexicelt, Lleisiwr, chwynnu 53 prompt)
5. estyn (*fetch*) y corpws
6. estyn y geiriadur ynganu
7. creu 10 set i brofi gyda'r dull C-V (isod) drwy ddefnyddio'r ffeiliau .csv;
8. Creu model iaith o'r corpws sy'n cael ei ddefnyddio yn y prawf;

Profi 10-plyg C-V

9. hyfforddi model acwstig ar sail y data sain;
10. profi'r modelau x10 gwaith:
    a. set hyfforddi + set brofi;
    b. llwytho un set hyfforddi a set brofi i HTK/*Kaldi* i gael y canlyniad;
    c. nodi WER% (isod);
11. cyfrifo cyfartaledd/gwyriad safonol (*standard deviation*)/cyfeiliornad safonol (*standard error*)/amrywiant (*variance*) (isod);
12. ailadrodd gyda fersiwn newydd o gorpws Paldaruo.

Gwelir o'r rhestr uchod bod angen paratoi'r data yn gyntaf i sicrhau bod y signal sain yn y fformat cywir a chreu ffeil .csv, sy'n cyfeirio at leoliad pob ffeil .wav (*waveform*) ar y cyfrifiadur. Gwneir hyn ar gyfer bob fersiwn Paldaruo er mwyn gallu eu rhannu a defnyddio'r dull C-V yn haws (isod). Mae hyn yn arwain at hyfforddi'r modelau ac yna, gwneud y profion C-V er mwyn cyfrifo canlyniad. Fel nodir uchod, er mwyn adeiladu modelau cadarn, rhaid eu hyfforddi ar set gynhwysfawr o frawddegau a geiriau. Yn ddelfrydol, dylai'r rhain fod wedi'u cydbwyso'n seinegol (Young, et al., 2015, tt. 26-28).

Mae'r geiriadur ynganu yn rhestr o'r geiriau yn y lecsicon gyda'u trawsgrifiadau cyfatebol. Ceir enghraifft o set o eiriau penodol yn nisgrifiad yr *HTK Book* o eiriadur ynganu ar gyfer tasg adnabod promptiau ffôn yn Saesneg (Young, et al., 2015, t. 28):

(17) a. CALL k ao l
b. DIAL d ay ax l
ch. EIGHT ey t
d. PHONE f ow n

Mae angen geiriadur ynganu cynhwysfawr i ddangos trawsgrifiadau geiriau yn y lecsicon. Ym mhrofion y traethawd hir hwn, defnyddir geiriadur ynganu (Chan & Cooper, 2014). Dyma enghraifft o'r cynnwys, lle mae'r priflythrennau yn dynodi llafariaid a'r llythrennau bach yn dynodi cytseiniaid:

(18) a. ABERAERON A b E r AE r O n
b. ABERAF A b E r A f
c. ABERAFAN A b E r A f A n
ch. ABERAI A b E r AI

Adeiladwaith Citiau Offer HTK a Kaldi

O fewn citiau offer HTK a *Kaldi*, mae dau fath o fodel yn cael eu hyfforddi sef model acwstig a model iaith. Yn syml, mae model acwstig yn cyfrifo tebygolrwydd cyfres o seiniau lle mae model iaith yn cyfrifo tebygolrwydd geiriau (Cooper, 2019, t. 1789). Mae model acwstig yn gynrychiolaeth ystadegol o nodweddion acwstig. Wrth echdynnu nodweddion acwstig y signal hwn, mae'n creu cynrychiolaeth ystadegol. Defnyddir dull

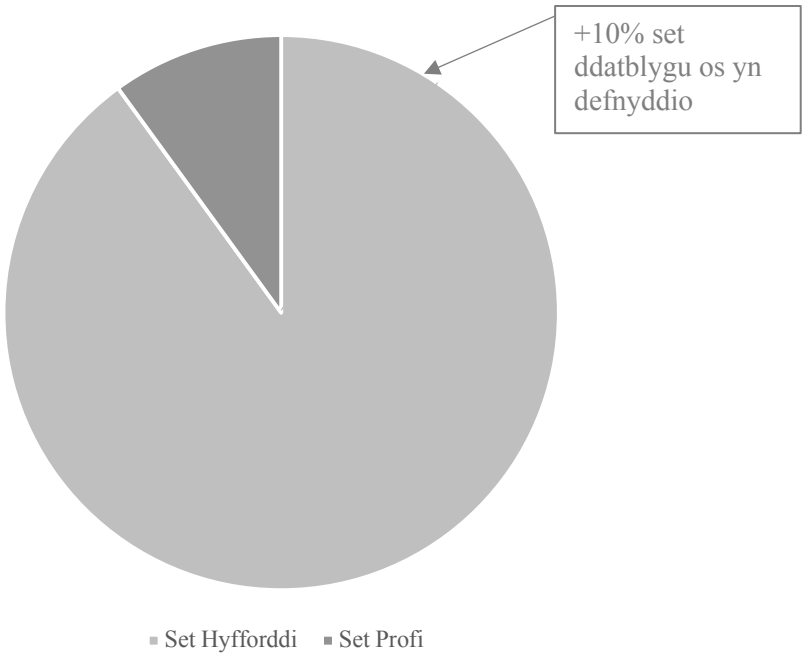
Mel-frequency cepstral coefficients (MFCC) i echdynnu'r nodweddion hyn (Young, et al., 2015, t. 33), (Povey, et al., 2011, t. 2). Mae'r dull hwn yn rhannu'r signal sain i ffenestri gwahanol er mwyn deall yr wybodaeth seinegol a dadansoddi'r sain (Gales & Young, 2007, t. 202). Mae modd modelu dosbarthiad y nodweddion hyn gyda GMM. I fodelu'r ffonemau gyda'u sain cyfatebol, defnyddir HMMs (Jones et al., 2019, t. 26).

Fel nodir uchod, mae'r profion hyn yn ffocysu ar fodelau acwstig felly caiff y modelau iaith eu creu o'r corpws sydd dan sylw ym mhob prawf unigol. Trafodir goblygiadau hyn ym mhennod VII Amrywio Iaith. Mae'r model iaith yn cyfrifo tebygolrwydd cyfres o eiriau (Jurafsky & Martin, 2019, t. 31). Gwneir hyn wrth ddadansoddi corpws testun, neu lecsicon, ac asesu pa mor aml mae un gair yn dod ar ôl n-gair. Wrth gyfuno'r wybodaeth hyn gyda gwybodaeth y model acwstig, mae modd cyfrifo tebygolrwydd cynnwys y signal sain drwy ddefnyddio'r hafaliad Bayes fel gwelir ym mhennod II 3 Modelau.

Fel nodir uchod yn 1.2, y cam ychwanegol i git offer *Kaldi* yw ychwanegiad y dull hyfforddi DNN. Dyma beth sy'n golygu bod *Kaldi* yn git offer croesryw yn defnyddio dull HMM-DNN yn hytrach nac HMM-GMM yn unig. Fel nodir ym mhennod II 3 Modelau, mae DNN yn fath o rwydwaith niwral sy'n ceisio efelychu niwronau'r ymennydd yn ysgogi un ar ôl y llall, yn hidlo gwybodaeth nes cyrraedd yr ateb cywir (Jones et al., 2019, t. 8). Mae DNN yn creu haenau cudd sy'n darganfod yr opsiwn mwyaf tebygol bob tro, yna'n symud ymlaen at yr haen nesaf i ailadrodd y broses. Mantais y dull hwn yw bod modd i'r system weithio dan lai o oruchwyliaeth yn y pendraw (Benigo, 2009). Dengys gwaith Gaida et al. (2014) bod yr hyfforddi ychwanegol hwn yn fanteisiol wrth adeiladu modelau acwstig Saesneg ac Almaeneg.

<u>Dulliau Profi</u>

Y ffordd orau o asesu perfformiad modelau yn ôl Jurafsky & Martin (2019, t. 36) yw asesiad anghynhenid (*extrinsic evaluation*) lle mae angen rhedeg y system gyfan eto bob tro mae unrhyw newid er mwyn mesur yr effaith. Mae'r fath asesiad yn broses ddrud ac araf lle mae angen rhedeg y system gyfan dro ar ôl tro. Prawf mwy ymarferol yw asesiad cynhenid (*intrinsic evaluation*) lle mae modd profi effeithiolrwydd y modelau gyda set brofi (Jurafsky & Martin, 2019, t. 36) fel gwelir isod yn Ffigwr 17.



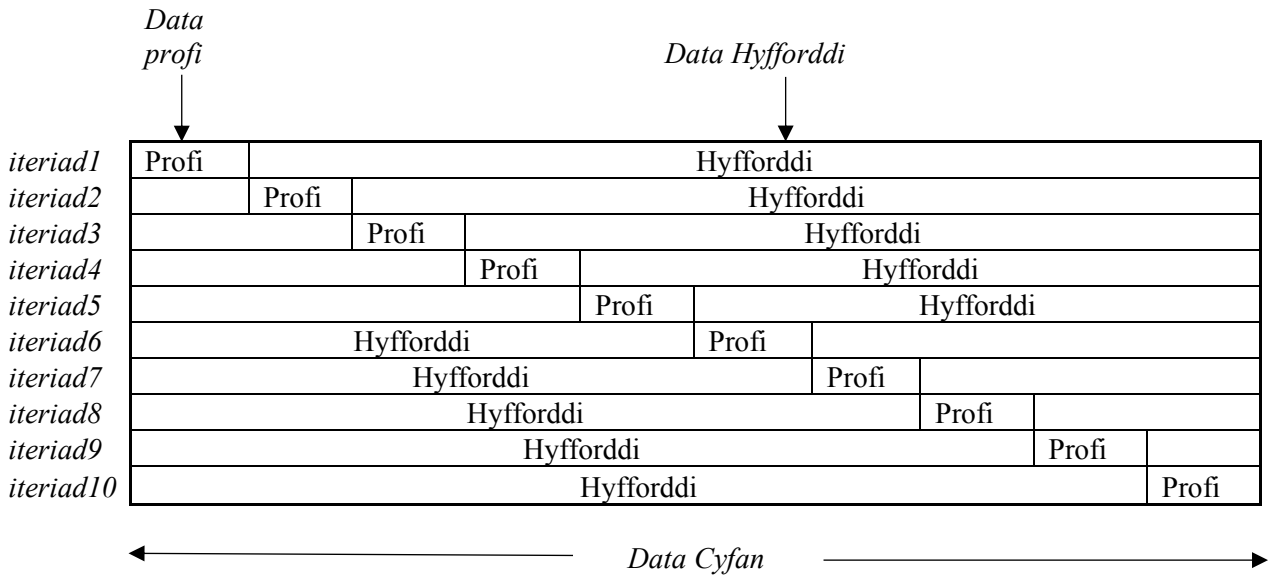
Ffigwr 17 Asesiad cynhenid

Wrth rannu'r corpws hyfforddi fewn i set hyfforddi a set brofi, mae modd asesu pa mor effeithiol yw'r modelau wrth gyfrifo tebygolrwydd cynnwys y signal sain. Y rhaniad arferol ar gyfer y setiau hyn yw: 90% hyfforddi a 10% profi (Jurafsky & Martin, 2019, tt. 36-37) fel gwelir yn Ffigwr 17 uchod. Wrth wneud hyn, mae'n bwysig peidio cynnwys data profi yn y set hyfforddi gan ei fod yn creu tuedd yn y canlyniadau. Mae hyfforddi ar y set brofi (*training on the test set*), yn gallu dangos tebygolrwyddau'n rhy uchel (Jones & Cooper, 2016, t. 76), (Cawley & Talbot, 2010, t. 2080). Hefyd, mae gorddefnydd o set brofi penodol yn gallu arwain at duedd tebyg, gorffitio (*overfitting*, (Cawley & Talbot, 2010)). Mae dwy ffordd o leihau'r tebygolrwydd o orffitio. Y tuedd gydag ieithoedd sydd â digon o adnoddau yw cyflwyno set ddatblygu, fel gwelir yn Ffigwr 17. Mae hwn yn set brofi newydd sbon, sydd wedi'i neilltuo tan ddiwedd y prawf (Jurafsky & Martin, 2019, tt. 36-37). Mae'r set ddatblygu fel arfer yn 10% o'r holl ddata (Ffigwr 17). Ffordd arall o leihau'r tebygolrwydd o orffitio yw defnyddio dull C-V (Ffigwr 18).

Er bod y dull cynhenid (Ffigwr 17) yn cynnig manteision o ran cost a chyflymder profi, mae diffyg adnoddau (diffyg data) yn achosi problemau pellach (Jones & Cooper, 2016, t. 76). Er mwyn defnyddio'r dull cynhenid, rhaid i'r set hyfforddi fod y mwyaf posib (Jurafsky & Martin, 2019, t. 36) ond, wrth orfod neilltuo'r data hwn ar gyfer profi (a datblygu os yn defnyddio), mae'n bosib na fydd digon o ddata i hyfforddi'n effeithiol

(2019, t. 69). Mae hyn yn arwain at duedd (Cawley & Talbot, 2010, t. 2102) ac er mwyn dilyn arfer da, dylid sicrhau bod canlyniadau'n ddiduedd (2010, t. 2080). Mae tuedd yn y cyd-destun hwn yn cyfeirio at sefyllfa lle nad yw'r canlyniadau'n adlewyrchu gwir allu'r system.

Mae C-V yn ddull i sicrhau nad oes tuedd wrth gyfrifo gwir allu'r system (2010, t. 2086), sy'n broblem gyffredin wrth ymdopi gyda data llai oherwydd y pwysau i neilltuo data ar gyfer profi. Fel nodir ym mhennod II 4 Dulliau i Ymdopi gyda Llai o Ddata, defnyddir y dull yn aml gan y gymuned dysgu peirianyddol (Jurafsky & Martin, 2019, t. 69), (Cawley & Talbot, 2010, t. 2082) gan gynnwys Caon et al. (2011) a welodd gwelliant mewn system Ffrangeg. Wrth ddefnyddio'r dull hwn (Ffigwr 18), mae angen rhannu'r data fewn i 10 *fold* (plyg) ar hap ac yna, rhannu bob plyg yn 90% (set hyfforddi) a 10% (set brofi). Mae angen rhedeg 10 prawf ar wahân i greu 10 canlyniad unigol. Wrth wneud hyn, mae modd cyfrifo cyfartaledd y 10 canlyniad. Mae hwn yn creu amcangyfrif mwy dibynadwy o effeithiolrwydd y cit offer nac wrth redeg un prawf yn unig a phrofi'r data cyfan ar ei ben ei hun fel sy'n digwydd wrth hyfforddi ar y set brofi (Cawley & Talbot, 2010, t. 2082). I greu'r data 10-plyg rhaid rhannu'r setiau mewn ffeil .csv cyn dechrau profi er mwyn creu'r 10-plyg.



Ffigwr 18 10-*fold cross-validation*

Oherwydd nad oes llawer o ddata, ac felly nad oes modd neilltuo 10% ychwanegol ar gyfer set ddatblygu, defnyddir y dull hwn er mwyn cymharu effeithiolrwydd y citiau offer HTK a *Kaldi* wrth ymdopi â data llai yn y traethawd hir hwn.

Cyfradd Geiriau Gwallus

Mae *Kaldi* yn defnyddio WER (cyfradd geiriau gwallus, (Povey, et al., 2011, t. 3)) ar gyfer cyfrifo canlyniadau. Gwelir y broses hon yn Hafaliad 2 isod. Dyma'r hafaliad i gyfrifo WER lle mae N yn cynrychioli nifer y geiriau caiff eu hadnabod yn y mewnbwn:

$$WER = \frac{I + D + S}{N}$$

Hafaliad 2 Cyfradd geiriau gwallus

Mae WER yn cyfrifo faint o wallau ar lefel geiriau sydd wedi digwydd wrth hyfforddi a phrofi. Felly, os yw'r WER yn uchel, mae'r system wedi darganfod nifer fawr o wallau, sy'n ganlyniad gwael, lle os yw'r WER yn isel, mae llai o wallau ac felly yn ganlyniad gwell. Mae'r cyfraddau hyn yn cyfrif 3 math gwahanol o wall fel gwelir isod (Young, et al., 2015, t. 232).

- mewnosodiad (*insertion*, I): pan mae gair wedi ei ychwanegu i'r allbwn nad oedd yn ymddangos yn y mewnbwn;
- dilead (*deletion*, D): pan mae gair oedd yn ymddangos yn y mewnbwn wedi ei ddileu o'r allbwn;
- amnewidiad (*substitution*, S): pan mae gair gwahanol yn yr allbwn i gymharu â'r rhan gyfatebol o'r mewnbwn.

Ar y llaw arall, mae HTK yn defnyddio W.Acc (cyfradd geiriau cywir, (Young, et al., 2015, t. 231)), sy'n cael ei dangos yn Hafaliad 3 isod o Young et al. (2015, t. 231).

$$WA.cc = \frac{N - I - D - S}{N} \times 100\%$$

Hafaliad 3 Cyfradd geiriau cywir

WER yw'r mesur gaiff ei ddefnyddio'n fwyaf cyffredin er mwyn mesur perfformiad system adnabod lleferydd (Jurafsky & Martin, 2019, t. 25). Bydd canlyniadau HTK yn cael eu trosi i WER er mwyn hwyluso'r gymhariaeth rhyngddynt. Gwelir y broses trosi i WER yn Hafaliad 4 isod.

$$\%WAcc = 100 - \%WER$$

Hafaliad 4 Trosi W.Acc i WER

Trwy gydol y traethawd hir hwn, defnyddir y gyfradd WER i gyfrifo canlyniadau profion. Cyfrifir ystadegau disgrifiadol hefyd er mwyn darganfod unrhyw ganlyniadau annisgwyl yn y dull C-V: gwyriad safonol, cyfeiliornad safonol ac amrywiant. Gwneir hyn i fesur gwasgariad y data o'r cyfartaledd neu'r cymedr. Wrth brofi gyda'r dull C-V, mae'n bwysig sicrhau nad yw'r 10 canlyniad wedi'u gwasgaru o'r cymedr yn ormodol.

3. Canlyniadau

Yn y rhan hon, ceir canlyniadau WER y profion yn gyntaf, ac yna dadansoddiad pellach o ganlyniadau *Kaldi* wrth adnabod lleferydd data Paldaruo.

Canlyniadau WER

Mae'r prawf hwn yn edrych os yw modelau ystadegol HTK neu fodelau croesryw *Kaldi* yn rhoi canlyniadau gwell gyda data llai. Er mwyn ymchwilio i'r cwestiwn hwn, mae angen cymharu canlyniadau'r ddau git offer gyda data o gorpws Paldaruo. Gwelir canlyniadau'r profion hyn isod yn Tabl 11. Nodwyd yn 2.1 bod y profion hyn yn defnyddio corpws Paldaruo sy'n gorpws bach o ddata sain Cymraeg wedi'i gynllunio'n arbennig ar gyfer prosesu lleferydd. Ceir hefyd esboniad o fersiynau gwahanol y corpws yn Tabl 10 ac ehangiad ar ddata llai ym mhennod I 2.4 Diffyg Data, Amrywio Ieithyddol a'r Dulliau Perthnasol. Mae'r canlyniadau yn Tabl 11 yn edrych ar WER y ddau git offer (HTK a *Kaldi*) wrth gynyddu maint y data, o ddefnyddio gwahanol fersiynau corpws Paldaruo. Mae cynnydd o 2 awr rhwng bob fersiwn Paldaruo, gan ddefnyddio copi glân fersiwn 4, hynny yw wedi golygu, hidlo a chwynnu fel esboniwyd uchod yn 2.2 Gweithdrefn). Mae'r WER yn cyfrifo faint o wallau sydd yn yr allbwn o'i gymharu â'r mewnbwn felly'r isaf yw'r WER, y gorau oll yw'r canlyniad. Defnyddir canlyniadau tri3b (ystadegol, 1.2) ar gyfer *Kaldi* oherwydd ei fod yn rhoi canlyniadau gwell na nnet2 (niwral, 1.2). Ceir trafodaeth o oblygiadau hyn isod yn y rhan hon, o dan y pennawd *Dadansoddiad Pellach.*

Fersiwn Paldaruo	2.0		3.0		4.0	
Oriau	34.0		36.0		38.0	
Cit Offer	HTK	*Kaldi* (trib3)	HTK	*Kaldi* (tri3b)	HTK	*Kaldi* (tri3b)
Cyfartaledd (WER%)	14.7	4.35	14.7	4.27	16.24	4.72
Gwyriad Safonol (N)	0.49	0.97	0.49	1.02	0.44	0.88
Cyfeiliornad Safonol (N)	0.16	0.31	0.15	0.32	0.14	0.28
Amrywiant (N)	0.24	0.94	0.24	1.04	0.19	0.77

Tabl 11 Cymharu citiau offer HTK a *Kaldi* (data corpws Paldaruo)

Wrth hyfforddi ar ddata Cymraeg, mae'r canlyniadau cyfartaledd yn dangos gwahaniaeth rhwng HTK a *Kaldi* (tua 10% WER). Cynhelir cymhariaeth debyg gan Povey et al. (2011, t. 3) a ddangosodd nad oedd gwahaniaeth mawr (tua 1% WER) rhwng HTK a *Kaldi* wrth hyfforddi ar ddata Saesneg. Wrth gymharu HTK a *Kaldi* yn y profion hyn (Tabl 11), mae *Kaldi* yn dangos WER o 10.77% yn is nac HTK ar gyfartaledd. Nodir nad yw hyn yn cyfateb gydag arwyddocâd ystadgeol (fel nodir uchod yn 2.2 Gweithdrefn, cyfrifwyd ystadegau disgrifiadol i ychwanegu at ddilysrwydd y canlyniadau). Dros yr holl fersiynau, mae gwyriad a chyfeiliornad safonol HTK a *Kaldi* yn eithaf tebyg sy'n dangos bod y canlyniadau yn gyson rhwng y plygiadau, hynny yw, heb ganlyniadau annisgwyl.

Wrth gymharu canlyniadau fersiynau Paldaruo, does dim gwahaniaeth yng nghanlyniadau HTK wrth ychwanegu cyn lleied â 2 awr ychwanegol o ddata rhwng fersiwn 2.0 a 3.0. Er mai dim ond ychydig o wahaniaeth sydd yn *Kaldi*, mae'r WER yn newid wrth ychwanegu data rhwng fersiwn 2.0 a 3.0 (gwelliant o 0.08% WER). Mae angen ychwanegu 2 awr ychwanegol o ddata ar ben hynny gyda fersiwn 4.0 cyn gweld newid yng nghyfartaledd HTK (gwaethygu o 1.54% WER). Ymddengys bod *Kaldi* yn fwy sensitif i newid ym maint y data nac HTK, wrth i'r cyfartaledd amrywio wrth ychwanegu cyn lleied â 2 awr o ddata. Dylid nodi bod canlyniadau *Kaldi* yn gwaethygu ychydig hefyd wrth ddefnyddio data fersiwn 4.0 Paldaruo (gwaethygu 0.45% WER). Mae'n bosib bod hyn yn awgrymu bod y data wedi cyrraedd pwynt lle nad oes modd gwella'r canlyniadau'n fawr wrth ychwanegu mwy o'r un data. I atgoffa, mae'r fersiynau dan sylw (Tabl 10) yn cyfeirio at fwy o'r un data yn unig, heb amrywio'r promptiau.

Dadansoddiad Pellach

Nodir uchod yn 1.2 bod *Kaldi* yn git offer sydd ag adeiladwaith croesryw, sy'n meddwl ei fod yn gymysgedd o fodelau ystadegol HMM-GMM a niwral DNN. I atgoffa, mae HTK yn defnyddio modelau ystadegol HMM-GMM y unig. Mae *Kaldi* yn creu dau set o ganlyniadau sy'n berthnasol i'r traethawd hir hwn: tri3b a nnet2. Er gwybodaeth, mae *Kaldi* yn rhoi 6 prif ganlyniad sy'n amrywio o mono, sy'n defnyddio monoffonau yn unig, i tri3b, sef trydydd iteriad y canlyniadau yn defnyddio triffonau, i nnet2, sy'n defnyddio modelau niwral. Y prif wahaniaeth rhwng canlyniadau tri3b a nnet2 *Kaldi* yw bod tri3b yn defnyddio modelau HMM-GMM, lle mae nnet2 yn defnyddio modelau DNN, sef dull gwahanol, mwy diweddar, o fodelu fel gwelir yn 1.2. Gwelir yn Tabl 11 mai canlyniadau tri3b sy'n cael eu cofnodi ar gyfer *Kaldi*, sef y canlyniadau gorau HMM-GMM o fewn y profion hyn ar gyfer *Kaldi*. Isod, gwelir y rhain, yn ogystal â

chanlyniadau HTK, yn erbyn canlyniadau nnet2 *Kaldi*. Unwaith eto, mae'r canlyniadau nnet2 yn cyfeirio at ganlyniadau *Kaldi* lle defnyddiwyd y modelau DNN.

Fersiwn Paldaruo	2.0			3.0			4.0		
Oriau	34			36			38		
Cit Offer	HTK	*Kaldi* (tri3b)	***Kaldi* (nnet2)**	HTK	*Kaldi* (tri3b)	***Kaldi* (nnet2)**	HTK	*Kaldi* (tri3b)	***Kaldi* (nnet)**
Cyfartaledd (WER%)	14.7	4.35	**8.87**	14.7	4.27	**8.78**	16.24	4.72	**9.42**
Gwyriad Safonol (N)	0.49	0.97	**0.87**	0.49	1.02	**0.95**	0.44	0.88	**0.6**
Cyfeiliornad Safonol (N)	0.16	0.31	**0.27**	0.15	0.32	**0.3**	0.14	0.28	**0.19**
Amrywiant (N)	0.24	0.94	**0.75**	0.24	1.04	**0.09**	0.19	0.77	**0.36**

Tabl 12 Dadansoddiad canlyniadau tri3b a nnet2 *Kaldi*

Gwelir yn Tabl 12 bod canlyniadau nnet2 *Kaldi* ychydig yn waeth na chanlyniadau tri3b *Kaldi* (gwahaniaeth o 4.58 rhwng cyfartaledd canlyniadau tri3b a nnet2 *Kaldi*, nnet2 yn dangos y canlyniadau gwaeth). Dengys hyn mai'r ffordd orau o ddefnyddio cit offer *Kaldi* gyda data chorpws Paldaruo (Tabl 10) yw i ddefnyddio canlyniadau tri3b sef cydrannau HMM-GMM y cit offer. Mae hyn yn awgrymu bod angen mwy o ddata eto cyn gallu gwneud yn fawr o dechnegau DNN *Kaldi* (canlyniadau nnet2). Wrth gymharu hyn gyda chanlyniadau HTK (sy'n git offer HMM-GMM llwyr), mae canlyniadau *Kaldi* dal yn gyffredinol yn well. Gwelir gwelliant o 10.77% WER wrth ddefnyddio *Kaldi* tri3b a gwelliant o 6.19% WER o ddefnyddio *Kaldi* nnet2 dros HTK, yn defnyddio corpws Paldaruo.

I grynhoi, mae dadansoddiad Tabl 12 yn dangos bod modelau croesryw *Kaldi* yn rhoi canlyniadau gwell na modelau ystadegol HTK wrth brosesu data Paldaruo. O fewn hynny, mae'r canlyniadau tri3b *Kaldi* yn cyfrifo canlyniadau gwell na nnet2 *Kaldi* sy'n dangos bod angen mwy o ddata eto er mwyn gwneud yn fawr o fodelau niwral y cit offer hwnnw. Gweler y rhan nesaf ar gyfer trafodaeth o hyn mae'r canlyniadau hyn yn awgrymu.

4. Trafodaeth

Yn y rhan hon, ceir trafodaeth o ganlyniadau 3 gan gynnwys canlyniadau WER HTK a *Kaldi* yn ogystal â dadansoddiad pellach o ganlyniadau tri3b a nnet2 *Kaldi*. Wrth edrych ar ganlyniadau Tabl 11 uchod, awgrymir bod *Kaldi* yn ymdopi'n well nac HTK wrth brosesu data Paldaruo (gwelliant o 10.77% WER wrth ddefnyddio *Kaldi*). Wrth gymharu meintiau gwahanol y corpws drwy ddefnyddio gwahanol fersiynau Paldaruo, amlygir bod *Kaldi* yn fwy sensitif i ymateb i ychwanegiadau data (dim newid yng nghanlyniadau HTK wrth brosesu Paldaruo 2.0 a 3.0). Gwelir hefyd bod y canlyniadau'n cyrraedd pwynt ar ôl fersiwn 3.0 lle nad yw'r canlyniadau'n gwella rhagor (gwaethygu o 1.54% WER gyda HTK, ac o 0.45% WER gyda *Kaldi*). Wrth ystyried mai corpws cyfun o bromptiau geiriau unigol yw Paldaruo, mae hwn yn codi'r cwestiwn a yw ychwanegu mwy o'r un data yn ymarfer da neu beidio. Codir y cwestiwn a fyddai ychwanegu data gwahanol, mwy amrywiol, yn helpu'r system i gyfrifo canlyniadau gwell. Caiff hyn ei drafod yn y bennod nesaf, VI Math o Ddata. Wrth ddadansoddi canlyniadau *Kaldi* ymhellach, ymddengys mai canlyniadau tri3b sy'n rhoi'r canlyniadau gwell o fewn *Kaldi*. Yn groes i'r disgwyl, mae hyn yn golygu mai cydrannau ystadegol *Kaldi* sy'n rhoi'r canlyniadau gorau yn hytrach na'r rhai niwral. Yn y rhan nesaf o'r bennod hon, ceir trafodaeth o ymchwil arall sydd wedi dod i ganlyniadau tebyg, gan gynnwys goblygiadau'r canfyddiad hwn ar gyfer y Gymraeg.

Fel nodir yn 3, mae canlyniadau tri3b *Kaldi* yn cyfeirio at y cydrannau ystadegol HMM-GMM yn yr adeiladwaith croesryw. Mae canlyniadau nnet2 yn cyfeirio at y cydrannau niwral ynddo (DNN). Mae'r modelau niwral yn ddull mwy diweddar o gysylltu'r sain â'r ffonemau. Er bod canlyniadau Tabl 12 yn dangos bod canlyniadau nnet2 *Kaldi* yn waeth na chanlyniadau tri3b *Kaldi*, dydy'r gwahaniaeth hwnnw ddim yn fawr (gwahaniaeth o 4.58% WER rhwng cyfartaledd *Kaldi* tri3b a chyfartaledd *Kaldi* nnet2). Mae'r gwahaniaeth tipyn yn fwy wrth gymharu canlyniadau tri3b *Kaldi* gyda chanlyniadau HTK (gwahaniaeth o 10.77% WER rhwng cyfartaledd *Kaldi* tri3b a chyfartaledd HTK) sy'n awgrymu mai *Kaldi* yw'r cit offer gwell yn y cyd-destun hwn. Mae'r ffaith bod canlyniadau tri3b *Kaldi* yn well na nnet2 yn awgrymu nad oes digon o ddata eto er mwyn gwneud yn fawr o gydrannau niwral *Kaldi*. Daethpwyd i'r un canlyniad mewn ymchwil i adnabod lleferydd Hindi, lle'r canlyniad oedd mai'r model tri3b oedd yn rhoi'r WER gorau mewn cymhariaeth o holl fodelau *Kaldi* (Sri et al., 2020), ac eto mewn ymchwil arall ar gyfer Hindi (Upadhyaya et al., 2017). O'r meini prawf yn

Tabl 8 cyn dechrau'r profion yn y bennod hon, *Kaldi* oedd i'w weld yn well na HTK o ran:

1. symlrwydd i'w ddefnyddio;
2. trwydded mwy agored;
3. defnyddio dulliau arloesol er mwyn gwneud yn fawr o ddata llai.

Wrth ystyried y canlyniadau sydd wedi eu cyflwyno uchod yn Tabl 11, mae'n bosib ychwanegu:

4. **canlyniadau WER gwell.**

Mae Tabl 11 yn cefnogi fy rhagdybiaeth gyntaf i raddau gan fod canlyniadau *Kaldi* yn well nac HTK wrth brosesu data Paldaruo. Mae tri3b *Kaldi* yn well na chanlyniadau HTK o 10.77% WER ar gyfartaledd. Ond, wrth ddadansoddi canlyniadau *Kaldi* ymhellach, ymddengys nad yw'n cefnogi fy rhagdybiaeth yn gyfan gan nad y cydrannau niwral sy'n rhoi'r canlyniadau gorau. Yn fy rhagdybiaeth gyntaf, rwy'n rhagdybio y bydd *Kaldi* yn well oherwydd ei fod yn gwneud yn fawr o ddatblygiadau diweddar rhwydweithiau niwral. Er bod *Kaldi* yn rhoi canlyniadau gwell nac HTK ym mhrofion Tabl 11, mae dadansoddiad Tabl 12 yn dangos mai modelau ystadegol *Kaldi* (tri3b) sy'n rhoi canlyniadau gwell na chanlyniadau niwral *Kaldi* (nnet2). Wrth gymharu canlyniadau fersiynau gwahanol Paldaruo, mae'r canlyniadau yn cefnogi fy ail ragdybiaeth hefyd, gan fod y canlyniadau yn cyrraedd pwynt lle nad ydynt yn gwella yn fawr rhagor wrth ychwanegu 2 awr mwy o'r un data (gwaethygu o 1.54% WER gyda HTK, ac o 0.45% WER gyda *Kaldi*). Fel nodir uchod, awgryma hyn bod angen ystyried amrywio'r math o ddata yn hytrach na dim ond ychwanegu mwy o'r un data dro ar ôl tro. Mae'r bennod nesaf, VI Math o Ddata, yn edrych i gyflwyno mathau gwahanol o bromptiau yn y data er mwyn gwella gallu system i adnabod lleferydd Cymraeg.

VI. Math o Ddata

1. Cyflwyniad

Mae'r bennod hon yn amlinellu'r profion a'r data a ddefnyddiwyd er mwyn ymchwilio cwestiwn ymchwil 2 (IV 5 Amcanion yr Ymchwil), ynghylch pa fath o ddata sy'n fwyaf effeithiol o fewn system adnabod lleferydd Cymraeg. Mae'n edrych i sut mae data ar ffurf geiriau unigol ac ar ffurf brawddegau llawn yn effeithio'r canlyniadau. Pwrpas y profion hyn yn benodol yw amrywio'r math o ddata, a darganfod pa fath sy'n fwy effeithiol pan fo data llai, fel sydd ar gyfer y Gymraeg. Yn y bennod hon, ceir esboniad o'r profion a gynhelir er mwyn profi ymarferoldeb y rhagdybiaethau perthnasol sydd i'w gweld yn 1.3. Fel y bennod flaenorol, mae'r bennod wedi'i rhannu i dair rhan yn dilyn y cyflwyniad: Dull (Data, Gweithdrefn), Canlyniadau a Thrafodaeth. Unwaith eto, bydd trafodaeth gyffredinol am oblygiadau ac arwyddocâd y canlyniadau ym mhennod IX Trafodaeth.

1.1. Yr Her: Math o ddata

Mae'r profion yn y bennod hon yn dilyn profion pennod V Citiau Offer lle codir y cwestiwn a yw amrywio math y data yn cael effaith gadarnhaol. Fel nodir ym mhennod II 2.1 Hanes LVCSR a Chynorthwywyr Deallus, mae LVCSR (adnabod lleferydd geirfa eang di-dor) yn her bwysig i'w goresgyn os yw system am adnabod lleferydd naturiol yn llwyddiannus. Mae hyn yn berthnasol yma gan ei fod yn ymwneud â ffenomenau cyfuno seiniau lle gall seiniau fynd drwy eilededdau sain mewn rhai cyd-destunau. Enghraifft o hyn yw cymathu (IV 3.3 Tafodiaith ac Acen), sy'n rhywbeth na all promptiau geiriau unigol ei gipio'n hawdd (esboniad isod). Mae system LVCSR yn cyfeirio at system sy'n gallu adnabod lleferydd di-dor, neu gysylltiedig; hynny yw, mwy na dim ond gorchmynion neu atebion ar ffurf geiriau unigol. Mae lleferydd di-dor yn fwy heriol na geiriau wedi'u hynysu oherwydd bod yn rhaid i'r system drosi'r lleferydd di-dor fewn i linyn o eiriau, ond heb i'r ffiniau gael eu hamlinellu'n glir gyda seibiant, fel byddai'n digwydd rhwng geiriau unigol (Cooper, 2019, t. 1788). Yn dechnegol, mae'n cyfeirio at y ffenestri neu'r ffiniau sy'n cael eu creu yn y broses adnabod lleferydd i benderfynu pa air sy'n cael ei glywed (Jurafsky & Martin, 2019, t. 169). Yn aml, ni chaiff y geiriau cyfagos eu hystyried wrth bennu ffiniau'r ffenestri yma, sy'n broblem gyda lleferydd di-dor gan fod geiriau cyfagos yn aml yn effeithio ar ynganiad geiriau eraill (enghraifft isod).

Wrth arbrofi gyda phromptiau ar ffurf brawddegau llawn, rhoddir mwy o gyd-destun ffonolegol i'r promptiau megis cymathu. Ceir cyflwyniad i hyn, ac eilededdau sain eraill, ym mhennod IV 3.3 Tafodiaith ac Acen. I'n hatgoffa, mae cymathu yn cyfeirio at seiniau sy'n newid wrth ymddangos wrth ochr sain arall. Ceir yr enghraifft isod yn Hannahs (2013, t. 50) lle mae'r /n/ yn newid i'r sain yr /m/ sydd wrth ei ochr:

(19) a. 'mae e'n mynd' [mai ɛm mɪnd]

Mae'n debygol na fyddai'r newid hwn yn cael ei adlewyrchu mewn promptiau geiriau unigol gan fod pob gair wedi'i rannu oddi wrth y llall.

b. 'mae' [mai]
c. 'ef' [ɛv]
ch. 'yn' [ən]
d. 'mynd' [mɪnd]

Wrth gyflwyno'r promptiau ar ffurf brawddegau llawn fodd bynnag, lle mae geiriau'n rhedeg mewn i'w gilydd, byddai gwell tebygolrwydd o gipio newidiadau fel hyn yn y data sain. Fel nodir uchod, mae hyn yn rhan bwysig o system LVCSR (adnabod lleferydd geirfa eang di-dor).

1.2. Cymharu Geiriau Unigol gyda Brawddegau Llawn

Fel nodir uchod yn 1.1, mae'r profion yn y bennod hon yn adeiladu ar brofion pennod V Citiau Offer. Mae'r ddwy bennod yn ffocysu ar ddiffyg data, gyda'r bennod hon yn canolbwyntio ar ba fath o ddata sydd fwyaf effeithiol. Fel nodir yng nghanlyniadau pennod V Citiau Offer, mae awgrym bod perygl cyrraedd pwynt lle nad oes modd gwella'r canlyniadau yn fawr, wrth ddim ond ychwanegu mwy o'r un math o ddata. Gwelir hyn yng nghanlyniadau Tabl 11 a Tabl 12, pennod V 3 Canlyniadau, lle mae'r WER yn gwaethygu ychydig wrth ychwanegu 2 awr mwy o ddata promptiau geiriau unigol yn unig, rhwng Paldaruo 3.0 a 4.0 yn *Kaldi*. Mae corpws Paldaruo yn gasgliad o bromptiau geiriau unigol ar y cyfan (gwelir pennod III 2.3 GALLU, Paldaruo a Macsen). Cynlluniwyd y data yn 2014 fel rhan o'r project GALLU yn Uned Technolegau Iaith, Canolfan Bedwyr, er mwyn datblygu promptiau geiriau unigol i ddal seiniau mwyaf cyffredin yr iaith.

Er mwyn osgoi'r pwynt sy'n amlygu rhwng fersiwn 3.0 a 4.0 Paldaruo, lle nad yw'r canlyniadau'n gwella rhagor yn *Kaldi*, codir y cwestiwn o arbrofi gyda mathau gwahanol o ddata. Ceir esboniad yn 2.1 o greu promptiau newydd ar ffurf brawddegau llawn. Mae'r profion yn y bennod hon wedi'u dylunio i gymharu ymarferoldeb y ddau fath gwahanol o bromptiau. Hynny yw, o ddeall effaith promptiau geiriau unigol o'r bennod flaenorol (V Citiau Offer), mae'r bennod hon yn ymchwilio i effaith ychwanegu promptiau brawddegau llawn ar ganlyniadau adnabod lleferydd Cymraeg.

1.3. Cwestiynau Ymchwil a Rhagdybiaethau Perthnasol

Fel nodir ym mhennod IV 5 Amcanion yr Ymchwil, y cwestiwn ymchwil dan sylw yn y bennod hon yw rhif 2 sy'n edrych ar ba fath o ddata sydd fwyaf effeithiol o fewn system adnabod lleferydd Cymraeg. I atgoffa, drwy gyflwyno amrywiadau Saesneg Indiaidd i gamau hyfforddi peiriant adnabod lleferydd Saesneg Prydeinig, mae canlyniadau Huang et al. (2020) yn awgrymu bod modd gwella canlyniadau adnabod lleferydd drwy gyflwyno mwy o enghreifftiau a chyd-destun perthnasol wrth hyfforddi. Fy rhagdybiaeth yw y bydd data ar ffurf brawddegau llawn yn rhoi mwy o gyd-destun ffonolegol (er enghraifft, eilededdau sain) i'r geiriau a'r seiniau yn y mewnbwn na data ar ffurf geiriau unigol yn unig, ac yn gwella effeithiolrwydd y system wrth gyfrifo tebygolrwydd y geiriau a'r seiniau mae'n eu hadnabod.

2. Dull

Bydd y rhan hon yn disgrifio'r dull o hyfforddi a phrofi modelau yn *Kaldi* gyda data brawddegau llawn, er mwyn cymharu canlyniadau pennod V Citiau Offer sy'n defnyddio fersiwn 4.0 corpws Paldaruo (geiriau unigol yn unig). Yn 2.1, ceir cyflwyniad byr i'r data a ddefnyddir a'r broses o greu promptiau ar ffurf brawddegau llawn. Er mwyn esbonio'r broses casglu data, bydd cyflwyniad i'r platfform *Common Voice-cy* (III 2.4 Mozilla*, Common Voice* a *Common Voice* Cymraeg (Mozilla, 2020)). Mae *Common Voice-cy* yn cynnig ehangiad o ymgyrch casglu data Mozilla, i ieithoedd llai eu hadnoddau fel y Gymraeg (cyflwyniad i'r term ym mhennod I 2.2 'Iaith Lai ei Hadnoddau'). Yn 2.2, esbonnir sut rhedir y profion gyda'r data hyn.

2.1. Data

Fel nodir ym mhennod V 2 Dull, mae corpws Paldaruo yn cynnwys promptiau geiriau unigol sy'n seinegol gytbwys, sy'n golygu ei fod yn adlewyrchu defnydd mwyaf cyffredin seiniau'r iaith Gymraeg (Prys & Jones, 2018). Er mwyn cynnal prawf teg â data newydd ar ffurf brawddegau llawn, gwneir ymdrech i adlewyrchu hyn yn y promptiau newydd. O dan y penawdau isod, ceir manylion data Paldaruo, proses ddatblygu promptiau newydd i gyflawni'r uchod, a manylion data *Common Voice-cy.*

Data Corpws Paldaruo

Yn y bennod hon, y ffocws yw math y data o fewn y fersiynau, felly cyflwynir manylion fersiynau Paldaruo yn Tabl 13, gydag ychwanegiad nifer y promptiau a math y data. Defnyddir cit offer *Kaldi* ar gyfer y profion hyn.

Fersiwn Paldaruo	2.0	3.0	4.0
Cyhoeddwyd	15/07/2016	09/06/17	06/11/17
Amser Sain (awr)	34	36	38
Nifer y promptiau (awr)	43	43	85
Math y Data	Geiriau unigol yn unig	Geiriau unigol yn unig	Geiriau unigol yn unig
Nifer o Ffeiliau	11,556	12,024	12,682
Nifer o Gyfranwyr	487	506	536

Tabl 13 Manylion promptiau Paldaruo

Does dim promptiau ychwanegol rhwng fersiynau 1.0 a 3.0 ond mae mwy o recordiadau o'r promptiau yna, sy'n gwella'r WER ychydig, o 0.08% WER (Tabl 11, V 3 Canlyniadau). I atgoffa, dyma un enghraifft o'r promptiau:

(20) a. LLEUAD MELYN AELODAU SIARAD FFORDD YMLAEN CEFNOGAETH HELEN

b. RHYBUDDIO, ELEN, UWCHRADDIO, HWNNW, BEIC, CYMRU, RHOI, AELOD

Erbyn fersiwn 4.0, ychwanegwyd promptiau newydd (42 prompt newydd) gan yr Uned Technolegau Iaith yn yr un ffurf: geiriau unigol. Nodir bod ychwanegu promptiau newydd, ond ar yr un ffurf o eiriau unigol, yn gwaethygu'r WER o 0.45% WER (Tabl 11, pennod V 3 Canlyniadau). Nodir mai dim ond 2 awr o ddata ychwanegol oedd yn wahanol rhwng fersiynau 3.0 a 4.0. Dyma enghraifft o'r promptiau newydd yn fersiwn 4.0:

(21) a. RHAWN UNGELLOG CHWITFFORDD DEHEUBARTH ROBERTS THAW HAWYS DDUW
b. GOWT JIWDO NIWCLEWS CORSIOG MEUDWYOL EINIOES BWÂU

Roedd Paldaruo dal yn casglu data yn ystod y cyfnod profi ac fe gyhoeddwyd fersiwn 5.0 yn 2018 oedd â 3,040 prompt newydd, gan gynnwys brawddegau llawn (hynny yw, yn cynrychioli lleferydd cysylltiedig heb seibiant rhwng geiriau unigol). Canolbwyntiai'r profion hyn ar *Common Voice-cy* wrth arbrofi gyda brawddegau llawn, gan ei fod wedi creu a chasglu adnoddau sylweddol mewn amser byr. Mae'r rhain yn cynnwys brawddegau greais i (manylion ac enghreifftiau isod) a phromptiau eraill gan yr Uned Technolegau Iaith gan gynnwys promptiau o'r project Macsen (III 2.3 GALLU, Paldaruo a Macsen), er enghraifft:

(22) a. BETH FYDD Y TYWYDD FORY
b. BETH OEDD NEWYDDION DDOE

Fel nodir isod wrth fanylu ar ddata *Common Voice-cy*, mae promptiau Paldaruo 5.0 yn cyfateb i bromptiau y fersiwn cynnar a ddefnyddiais o *Common Voice-cy* (13/02/19). Gwelir holl bromptiau Paldaruo 1.0 i 4.0 yn Atodiad A ac Atodiad B. Mae sampl o bromptiau Paldaruo 5.0 yn ymddangos yn Atodiad C fel rhan o sampl data fersiwn cynnar *Common Voice-cy* (13/02/19).

Datblygu Promptiau Newydd

Fel nodwyd uchod, mae promptiau Paldauro wedi'u cynllunio i fod yn seinegol gytbwys ac felly'n adlewyrchu seiniau mwyaf cyffredin yr iaith Gymraeg. I ddatblygu promptiau newydd ar ffurf brawddegau llawn sy'n cyfateb yn y modd hwn, defnyddir trawsgrifiadau corpws o ddata sgyrsiol, *Siarad* Bangor (Deuchar, 2014). Defnyddir y corpws *Siarad* Bangor er mwyn ceisio adlewyrchu'r geiriau mwyaf cyffredin (yn codi'n amlaf yn y corpws hynny yw) yn y brawddegau newydd.

Pwrpas casglu corpws *Siarad* Bangor oedd cipio enghreifftiau naturiol o gyfnewid cod (IV 3.2 Cyfnewid Cod) rhwng siaradwyr Cymraeg (Deuchar, 2014, t. 1). Er mwyn adlewyrchu promptiau Paldaruo (Cooper et al., 2019, tt. 3-6), mae angen hidlo'r corpws am amrywiadau cyfnewid cod a chywair (IV 3.1 Cywair; ymchwil pellach ym mhennod VII Amrywio Iaith). Defnyddir diffiniadau ISO (ISO, 2018) ar gyfer diffinio cyweiriau

gwahanol yn y traethawd hir hwn. Yna, mae angen trefnu corpws *Siarad* Bangor yn ôl amlder geiriau. Defnyddir meddalwedd CLAN i wneud hyn, er mwyn cipio'r geiriau mwyaf cyffredin. Ceir rhai o'r enghreifftiau mwyaf cyffredin isod:

(23) I, YN, Y, AR, AM, DDIM, O, MAE, TI, MYND

Mae'r corpws yn cynnwys tua 40 awr o ddata gan 151 cyfrannwr ac wedi'i drwyddedu dan drwydded agored sy'n meddwl bod modd ei ddefnyddio am unrhyw bwrpas (Deuchar, 2014, t. 1). Defnyddir amlder geiriau mwyaf cyffredin y data hwn felly er mwyn creu brawddegau llawn wedi seilio ar ddata corpws *Siarad* Bangor, sy'n rhoi rhywfaint o argraff o eiriau cyffredin yr iaith Gymraeg sgyrsiol (Deuchar, 2014, t. 1). Dyma enghreifftiau o rai o'r brawddegau newydd a grëir o fewn y traethawd hir hwn, ar sail data *Siarad* Bangor:

(24) a. DYDW I DDIM YN BWRIADU BOD YNG NGHAERDYDD DROS Y GWYLIAU.
b. MAE ANGEN I TI OFYN AM BETH HOFFET TI GAEL YN Y BWYTY.
c. OEDD RHAID I TI DDWEUD NAD OEDDET TI'N GWYBOD UNRHYWBETH?

<u>Data Common Voice</u>

Fel nodir ym mhennod III 2.4 Mozilla, *Common Voice* a *Common Voice* Cymraeg, erbyn mis Gorffennaf 2017, roedd cwmni technoleg Mozilla wedi cyhoeddi ymgyrch casglu data sain ar gyfer adnabod lleferydd Saesneg sef *Common Voice* (Ardila, et al., 2020, t. 4218). Ehangwyd y platfform hwn ar gyfer ieithoedd eraill ym mis Mehefin 2018, gan gynnwys y Gymraeg (Ardila, et al., 2020).

Cefais fynediad cynnar at gorpws *Common Voice-cy* ym mis Chwefror 2019, cyn cyhoeddi'r fersiwn cyntaf swyddogol ym mis Rhagfyr 2019. Roedd y fersiwn cynnar yn cynnwys promptiau Paldaruo 4.0, fy mrawddegau newydd i o'r dull uchod, a brawddegau newydd a ychwanegwyd gan yr Uned Technolegau Iaith (Atodiad C). Wrth i'r gronfa dyfu, roedd lleisiau newydd yn darllen y promptiau hyn o'r newydd hefyd. Gwelir yn Tabl 14, manylion corpws Paldaruo 4.0 a *Common Voice-cy* (13/02/19).

Data	Paldaruo 3.0	Paldaruo 4.0	*Common Voice-cy*
Cyhoeddwyd	09/06/17	06/11/17	13/02/19
Amser Sain (awr)	36	38	42
Nifer o Gyfranwyr	506	536	726

Tabl 14 Cymharu data Paldaruo a *Common Voice-cy*

Yn Tabl 11, pennod V 3 Canlyniadau, gwelir bod ychwanegu mwy o bromptiau geiriau unigol rhwng Paldaruo 3.0 a 4.0 yn gwaethygu'r WER ychydig (0.45% WER). Pwrpas y prawf yn y bennod hon yw edrych i weld os yw ychwanegu math gwahanol o bromptiau, brawddegau llawn, yn cael effaith gadarnhaol ar y canlyniadau (1.3). Er nad oedd y corpws ei hun yn gyhoeddus ar y pryd, roedd cynnyrch *Common Voice-cy* yn amlwg yn cyrraedd nifer mwy o gyfranwyr ac yn casglu maint mwy o oriau, felly roedd yn cynnig manteision dros ddefnyddio Paldaruo i gasglu data.

Yn ogystal â'r hyn a welir yn Tabl 14, mae canran o oriau corpws cyhoeddedig *Common Voice-cy* (cy_77h_2019-12-10) wedi'u dilysu. Ystyr hyn yw bod yr oriau hyn yn fersiwn 10/12/19 wedi'u gwirio gan gyfranwyr gyda chadarnhad bod y sain yn cyfateb i'r testun. Os nad yw'r sain yn ddibynadwy yn y modd hwn, caiff ei wrthod a'i ddiystyru o'r corpws cyhoeddedig.

I grynhoi'r is-ran Data, dyluniwyd promptiau newydd ar ffurf brawddegau llawn ac fe'u llwythwyd nhw i blatfform newydd, *Common Voice-cy*, oedd i'w weld yn cyrraedd nifer mwy o gyfranwyr ac yn casglu mwy o ddata.

2.2. Gweithdrefn

Fel nodir uchod, pwrpas y profion yn y bennod hon yw cymharu canlyniadau Paldaruo 4.0 (4.72% WER, Tabl 11, V 3 Canlyniadau), gyda chanlyniadau *Common Voice-cy* (13/02/19) er mwyn gweld beth yw effaith ychwanegu mwy o bromptiau newydd ond ar ffurf wahanol: brawddegau llawn. Gwelir y data gaiff ei ddefnyddio yn y prawf yn y bennod hon isod yn Tabl 15.

Data	Paldaruo 4.0	*Common Voice-cy* (fersiwn cynnar cyn cyhoeddi swyddogol)
Cyhoeddwyd	06/11/17	13/02/19
Amser Sain (awr)	38	42
Nifer o Gyfranwyr	536	726

Tabl 15 Data profi Paldaruo 4.0 a *Common Voice-cy*

Defnyddir y cit offer *Kaldi* gan y dangosodd hwnnw'r canlyniadau gorau, fel gwelir ym mhennod V 3 Canlyniadau wrth gymharu gyda chit offer. Gan fod data *Common Voice-cy* (13/02/19) yn dal yn gasgliad bach (42 awr), defnyddir y dull k-*fold cross-validation* (C-V) (II 4 Dulliau i Ymdopi gyda Llai o Ddata, Ffigwr 9) eto yn y profion hyn. I atgoffa, mae'r dull C-V yn cael ei ddefnyddio'n aml yn y gymuned dysgu peirianyddol (Jurafsky & Martin, 2019, t. 69). Mae'n ddull sy'n ddefnyddiol pan fo diffyg data gan nad oes angen neilltuo unrhyw ddata er mwyn profi ar wahân, fel sydd angen gwneud wrth ddefnyddio dulliau mwy traddodiadol megis y dulliau cynhenid (*instrinsic evaluation*) ac anghynhenid (*extrinsic evaluation*) (esboniad ym mhennod V 2.2 Gweithdrefn). Yn hytrach, mae modd defnyddio'r data cyfan wrth ei rannu i 10 plyg (*fold*) ar hap ac yna, rhannu bob plyg yn 90% set hyfforddi a 10% set brofi, yna rhedeg 10 prawf a chymryd cyfartaledd o'r canlyniadau. Wrth wneud hyn, mae modd gwneud yn fawr o'r data sydd ar gael, drwy leihau'r siawns am duedd yn y data gan nad yw'r un data yn cael ei ddefnyddio dro ar ôl tro. Mae ailddefnyddio data yn ormodol yn gallu arwain at orffitio sy'n mynd yn erbyn arfer da yn y maes dysgu peirianyddol (Cawley & Talbot, 2010, tt. 2102, 2080, 2086).

I grynhoi, dilynir camau pennod V 2.2 Gweithdrefn er mwyn trosi'r data i ffeiliau .csv a chreu 10 plyg, ac yna rhedeg cit offer *Kaldi* er mwyn creu'r modelau a chynnal y profion. Fel ym mhennod V 2.2 Gweithdrefn, rhedir 10 prawf a chyfrifir cyfartaledd yn ôl y dull C-V (II 4 Dulliau i Ymdopi gyda Llai o Ddata, Ffigwr 9). Defnyddir y metrig WER ar gyfer y canlyniadau ac fe geir esboniad o'r metrig hwn ym mhennod V 2.2 Gweithdrefn. Mae'r metrig yn cyfrif nifer y dileadau, mewnosodiadau ac amnewidiadau ar lefel gair, felly yr isaf yw'r WER, y gorau oll yw'r canlyniad. Fel yn y bennod honno,

V Citiau Offer, cyfrifir gwyriad safonol, cyfeiliornad safonol ac amrywiant er mwyn mesur gwasgariad y data o'r cyfartaledd neu'r cymedr. Fel nodir yna, mae hyn yn bwysig wrth ddefnyddio'r dull C-V er mwyn sicrhau nad oes canlyniadau annisgwyl rhwng y plygiadau (V 2.2 Gweithdrefn).

3. Canlyniadau

Yn dilyn y canfyddiad o bennod V Citiau Offer, sef bod perygl o gyrraedd pwynt lle nad oes modd gwella'r canlyniadau yn fawr wrth ddim ond ychwanegu mwy o'r un data, cynhelir prawf gan ychwanegu promptiau newydd ar ffurf brawddegau llawn. Gwneir hyn er mwyn darganfod beth yw effaith y canlynol ar ganlyniadau adnabod lleferydd Cymraeg:

1. promptiau geiriau unigol yn unig;
2. ychwanegu promptiau brawddegau llawn.

Gwelir canlyniadau'r profion hyn yn *Kaldi* isod yn Tabl 16. Mae'r canlyniadau yn edrych ar WER (2.2) mathau gwahanol o ddata (geiriau unigol ac yna ychwanegiad brawddegau llawn), gan ddefnyddio corpws Paldaruo fersiwn 4.0 a phromptiau *Common Voice-cy* (13/02/19) sy'n cynnwys promptiau newydd ar ffurf brawddegau llawn yn ogystal â geiriau unigol gwreiddiol Paldaruo.

	Math o bromptiau	Geiriau unigol	Geiriau unigol + brawddegau llawn
	Data	Paldaruo 4.0	*Common Voice-cy* (13/02/19)
	Oriau	38	42
WER%	**Cyfartaledd**	4.72	3.44
	Gwyriad Safonol	0.88	0.17
	Cyfeiliornad Safonol	0.28	0.06
	Amrywiant	0.77	0.03

Tabl 16 Cymharu promptiau geiriau unigol (Paldaruo 4.0) a brawddegau llawn (*Common Voice-cy* (13/02/19))

Mae Tabl 16 yn dangos canlyniadau *Kaldi* (tri3b) wrth ddefnyddio data geiriau unigol yn unig ac yna, eto gydag ychwanegiad data brawddegau llawn. Defnyddir data Paldaruo 4.0

gan mai dyna'r fersiwn mwyaf diweddar ar adeg profi, fel nodir ym mhennod V 2.1 Data. Daw canlyniadau Paldaruo 4.0 o Tabl 11, pennod V 3 Canlyniadau ond dyma'r tro cyntaf y defnyddir data *Common Voice-cy* yn y traethawd hir hwn. Wrth gymharu canlyniadau Tabl 16, gwelir bod ychwanegu brawddegau llawn yn gwella'r WER ychydig, o 1.28% WER. Awgryma hyn bod ychwanegu promptiau ar ffurf brawddegau llawn yn ddull ymarferol o wella'r canlyniad. Fodd bynnag, nid yw'r gwahaniaeth hwn yn fawr (nid yw gwahaniaeth o 4.72% i 3.44% yn fawr iawn yn y data a ddadansoddwyd). Beth sy'n dangos newid mwy, er dal ddim yn wahaniaeth mawr, yw mesuriadau'r gwyriad safonol. Mae gwyriad safonol canlyniadau *Common Voice-cy* 0.71 yn llai na chanlyniadau Paldaruo, sy'n awgrymu bod canlyniadau'r plygiadau yn dangos llai o wasgariad o'r cymedr (gwelir canlyniadau bob plyg yn Atodiad G). Wrth gymharu gyda data geiriau unigol yn unig felly, ymddengys bod data brawddegau llawn yn rhoi canlyniadau mwy cyson rhwng y plygiadau.

I grynhoi, gwelir yn Tabl 16 bod ychwanegu data ar ffurf brawddegau llawn yn rhoi canlyniadau ychydig yn well na data ar ffurf geiriau unigol yn unig wrth brosesu lleferydd Cymraeg. O fewn hynny, mae mesuriad y gwyriad safonol yn dangos bod brawddegau llawn yn rhoi canlyniadau sydd â llai o wasgariad o'r cymedr wrth ddefnyddio'r dull k-*fold cross-validation* (C-V) sy'n awgrymu canlyniadau mwy cyson. Er hyn, nid oes gwahaniaeth mawr i'w weld rhwng y canlyniadau.

4. Trafodaeth

Yn y rhan hon, ceir trafodaeth o ganlyniadau 3. Wrth edrych ar ganlyniadau Tabl 16 uchod, awgrymir bod data ar ffurf brawddegau llawn yn rhoi canlyniadau ychydig yn well na data ar ffurf geiriau unigol. Wrth edrych yn fanylach ar ganlyniadau Tabl 16, dim ond gwahaniaeth o 1.28% WER sydd i'w weld rhwng canlyniadau geiriau unigol a brawddegau llawn, sydd ddim yn newid mawr. Ymddengys fodd bynnag, bod ychwanegu promptiau ar ffurf brawddegau llawn felly, yn ffordd o wella'r canlyniadau ymhellach na'r pwynt lle nad yw'r canlyniadau'n gwella, fel gwelir yng nghanlyniadau pennod V 3 Canlyniadau. Wrth ychwanegu mwy o'r un math o ddata yn unig, gwelir gwaethygiad o 0.45% WER (wrth ddefnyddio canlyniadau fersiynau 3.0 a 4.0 Paldaruo gyda chit offer *Kaldi*) (V 3 Canlyniadau).

Mae hyn yn cefnogi gwaith Ball & Williams (2001, tt. 50, 54-55) sy'n nodi bod amrywiadau megis eilededdau sain yn amrywio rhwng y lleferydd cysylltiedig a geir mewn brawddeg lawn, a geiriau unigol. Mae'n adlewyrchu gwaith Cooper (2019, t. 1788) hefyd, sy'n nodi, yn fwy penodol i'r maes hwn, bod adnabod lleferydd cysylltiedig yn fwy heriol i brosesu gan fod angen prosesu llinyn o eiriau yn hytrach na geiriau wedi'u hynysu gan seibiant. I adeiladu ar hyn, mae'n codi'r cwestiwn o ba fathau eraill o amrywio gellid ystyried er mwyn gwella'r canlyniadau eto, a sicrhau bod mathau gwahanol o leferydd yn cael eu hadnabod yn llwyddiannus. Yn y bennod hon, dydy'r brawddegau newydd ddim yn cynnwys unrhyw amrywiadau mewn acen neu dafodiaith benodol (IV 3.3 Tafodiaith ac Acen). Gwneir hyn gan nad yw promptiau Paldaruo yn adlewyrchu'r fath amrywiadau (Cooper et al., 2019, tt. 3-6). Coda'r cwestiwn felly, wrth ychwanegu'r amrywiadau hyn i'r promptiau, a fyddai posibilrwydd o achosi newid mwy yn y canlyniadau. Mae'r posibilrwydd hwn yn arwain at y cwestiwn ymchwil nesaf ynghylch amrywio iaith. Coda'r cwestiwn a yw'r system adnabod lleferydd sydd wedi'i chreu cyn belled yn gallu ymdopi â ffurfiau gwahanol, gan nad oes enghreifftiau penodol wedi'u cynnwys yn y data hyfforddi. Wrth arbrofi gydag effaith amrywio pellach, byddai modd darganfod i ba raddau oedd angen i'r system ymdopi gydag amrywio iaith. Caiff y cwestiwn hwn a'i oblygiadau ei ymchwilio a'i drafod yn y bennod nesaf, VII Amrywio Iaith.

I grynhoi, dengys Tabl 16 bod fy nghanlyniadau yn cefnogi fy rhagdybiaeth i raddau gan bod ychwanegu data ar ffurf brawddegau llawn yn gwella'r canlyniadau. Mae ychwanegu data ar ffurf brawddegau llawn yn rhoi canlyniad gwell (1.28% WER) na data ar ffurf geiriau unigol yn unig, er dylid nodi nad yw'r gwahaniaeth hwn yn ddigon mawr i

ddangos gwelliant mawr. Fel nodir uchod, awgryma hyn bod rhywfaint o effaith i'w weld wrth amrywio math y data o eiriau unigol i frawddegau llawn, ond bod angen ystyried enghreifftiau eraill o amrywio rhwng siaradwyr fel data o ardaloedd gwahanol. Mae'r bennod nesaf, VII Amrywio Iaith, yn edrych i gyflwyno 2 set o ddata: de a gogledd, er mwyn gweld os yw'r system yn ymdopi gyda'r amrywio yn yr un ffordd. Caiff oblygiadau hyn eu trafod ymhellach ym mhennod IX Trafodaeth.

VII. Amrywio Iaith

1. Cyflwyniad

Mae'r bennod hon yn amlinellu'r profion a gynhelir gyda data acen benodol er mwyn ymchwilio cwestiwn ymchwil 3, ynghylch sut mae amrywio acen yn effeithio system adnabod lleferydd Cymraeg. Mae'n edrych i effaith amrywio iaith ar system adnabod lleferydd Cymraeg. Ceir cyflwyniad i amrywio iaith ym mhennod IV Cefndir: Amrywio yn y Gymraeg. Pwrpas y profion hyn yn benodol yw darganfod beth yw effaith hyfforddi a phrofi system adnabod lleferydd gan ddefnyddio data deheuol a data gogleddol (1.2). Yn y bennod hon, ceir esboniad o'r profion a gynhelir er mwyn profi ymarferoldeb y rhagdybiaethau perthnasol sydd i'w gweld yn 1.3. Mae'r bennod wedi'i rhannu yn yr un drefn a'r penodau blaenorol, sef i dair rhan yn dilyn y cyflwyniad: Dull (Data, Gweithdrefn: Profi, Gweithdrefn: Dadansoddi), Canlyniadau (Canlyniadau WER, Dadansoddi Gwallau) a Thrafodaeth. Bydd trafodaeth gyffredinol am oblygiadau ac arwyddocâd y canlyniadau ym mhennod IX Trafodaeth.

1.1. Yr Her: Amrywio Iaith

Mae'r profion yn y bennod hon yn ffocysu ar amrywio acen, yn benodol enghreifftiau seinegol megis:

(25) a. *ti* [ti:], *ti* [ti:] (De)
b. *ti* [ti:], *tŷ* [tɨ:] (Gogledd)

Wrth gyfeirio at acen yn y traethawd hir hwn, defnyddir diffiniad Rees (2016) wrth iddo wahaniaethu rhwng y termau 'tafodiaith' ac 'acen'. Yn ôl Rees (2016), mae 'acen' yn cyfeirio at nodweddion ffonolegol (trefn/patrwm y seiniau) a seinegol (cynhyrchiad sain) y siaradwr, tra bod 'tafodiaith' yn hytrach, yn cyfeirio at elfennau geirfaol a gramadegol (megis amrywio cystrawennol), yn ogystal â'r nodweddion ffonolegol a seinegol. Felly, gellir dweud bod 'acen' yn elfen o 'dafodiaith' ond nad yw pob ffurf wahanol 'tafodieithol' yn elfen o 'acen'.

Oherwydd bod acen yn amrywio mewn nodweddion seinegol, y model acwstig sydd dan sylw i weld os oes gwahaniaeth wrth ymateb i leferydd o'r de a lleferydd o'r gogledd (II 3 Modelau). I atgoffa, mae'r model acwstig yn gynrychiolaeth ystadegol o nodweddion acwstig ffonem neu gyfuniad o ffonemau, wedi'u creu o'r data hyfforddi yn

y profion hyn (Cooper, 2019, t. 1789). Mae'r system yn defnyddio'r model hwn, yn ogystal â'r geiriadur ynganu a'r model iaith, er mwyn darganfod y gair mwyaf tebygol. I atgoffa, mae'r geiriadur ynganu yn ffeil sy'n cynnwys pob gair yn yr eirfa gyda'i drawsgrifiad cyfatebol mewn ffonemau, sy'n dod o'r LTS (rheolau llythyren-i-sain) (Jurafsky & Martin, 2019, t. 518). Os yw'r system am adnabod ynganiadau gwahanol mewn acenion yn gywir, mae angen i'r ynganiadau gwahanol hyn gael eu cynnwys yn y data hyfforddi ac, yn ddelfrydol, y wybodaeth ieithyddol hon hefyd. Dydy amrywio hyd llafariaid er enghraifft, fel gwelir yn yr enghraifft uchod 'ti' a 'tŷ', heb eu cynnwys eto. Mae hyn oherwydd cyfyngiadau amser yn y project GALLU (III 2.3 GALLU, Paldaruo a Macsen) lle crëwyd yr adnodd ieithyddol. Er mwyn adnabod y ffurfiau gwahanol hyn yn effeithiol, rhaid comisiynu gwaith arbenigol i edrych i amrywio acenion gwahanol. Ar ôl rhedeg y profion, dengys dadansoddiad 3.1 bod angen edrych ar y model iaith hefyd er mwyn adnabod y ffurfiau gwahanol hyn yn effeithiol (trafodaeth bellach yn 4).

Mae'r profion yn y bennod hon wedi'u dylunio i edrych i sut mae system adnabod lleferydd Cymraeg lle nad yw manylion acen bob cyfrannwr wedi nodi, yn ymdopi gyda'r her benodol hon. Hynny yw, beth yw effaith profi system adnabod lleferydd ar amrywio acen fel yr enghraifft uchod, ond heb ei hyfforddi arno.

1.2. Cymharu De a Gogledd

Mae Awbery (1996, t. 13) yn nodi bod acenion yn amrywio rhwng siaradwyr yng Nghymru, fel nodir ym mhennod IV 3.3 Tafodiaith ac Acen. Mae Awbery yn adleisio gwaith Thomas & Thomas (1989, t. 32) wrth gyflwyno'r rhaniad rhwng y de a'r gogledd drwy ddefnyddio'r 'u-ogleddol'. Gwelir enghraifft o'r nodwedd hwn uchod yn 1.1, gan ddefnyddio'r geiriau 'ti' a 'tŷ'. Yn yr enghraifft hon, mae'r ynganiadau yn wahanol yn y gogledd, lle nad oes gwahaniaeth i'w glywed yn y de. Dylid pwysleisio bod hyd llafariaid eraill yn gallu amrywio rhwng y de a'r gogledd hefyd, fel nodir ym mhennod IV 3.3 Tafodiaith ac Acen. Gwelir yr enghreifftiau isod gan Awbery (1996, t. 14) a Thomas & Thomas (1989, tt. 34-35):

(26) a. *cosb* [cɔsb] (De); [co:sb] (Gogledd)
b. *pell* [pe:ɬ] (De); [pɛɬ] (Gogledd)

Ystyrir gwahaniaethau eraill yn y profion hyn megis colled seiniau:

(27) a. *haf* [haːf] (De); [haː] (Gogledd)

neu gyfnewid ‘e’ ac ‘a’:

b. *hanner* [hanɛr] (De); [hanar] (Gogledd)

Dydy’r ffin benodol rhwng y de a’r gogledd ddim yn amlwg, er bod rhai wedi rhoi cynnig arni (Rees, 2016; Awbery, 1996, t. 13; Thomas & Thomas, 1989, tt. 31-32). Fel nodir ym mhennod IV 3.3 Tafodiaith ac Acen, mae eithriadau ac amrywiaethau pellach, yn enwedig mewn rhanbarthau llai, sy’n cymylu’r ffin rhwng y de a’r gogledd. Fel nodir uchod, mae’r profion hyn yn edrych i amrywiadau seinegol de a gogledd Cymru, fel gwelir yn yr enghreifftiau uchod, ar system adnabod lleferydd Cymraeg.

1.3. Cwestiynau Ymchwil a Rhagdybiaethau Perthnasol

Mae’r profion yn y bennod hon yn ymchwilio i gwestiwn ymchwil 3 sy’n edrych i sut mae amrywio acen yn effeithio system adnabod lleferydd Cymraeg. Fy rhagdybiaeth gyntaf yw y bydd data acen benodol yn cael effaith mawr ar ganlyniadau adnabod lleferydd Cymraeg wrth hyfforddi â data Paldaruo 4.0 (V 2.1 Data), a data *Common Voice-cy* (13/02/19, VI 2.1 Data).

Gan bod y mwyafrif o gyfranwyr Paldaruo 4.0 o’r gogledd (51.45% o’r data yn dod o’r Gogledd, lle bod 28.87% o gyfranwyr o’r de, Tabl 17), fy ail ragdybiaeth yw bydd y set brofi gogleddol yn rhoi canlyniadau gwell na’r set brofi deheuol, gan fod mwy o ddata o’r gogledd yn y data hyfforddi (1.2).

Lleoliad Acen	**Nifer Cyfranwyr**	**% Paldaruo 4.0**
Canolbarth	75	14.42%
De-ddwyrain	75	14.45%
De-orllewin	101	19.46%
Gogledd-ddwyrain	52	9.83%
Gogledd-orllewin	217	41.62%
Cyfanswm	520	99.78%

Tabl 17 Metadata lleoliad acen cyfranwyr corpws Paldauro (fersiwn 4.0) - peth metadata ar goll

Nid oes rhaid cofrestru manylion fel lleoliad acen er mwyn cyfrannu at *Common Voice-cy* felly mae'n anodd rhagdybio os byddai'n ymdopi yn well na Paldaruo 4.0 ar sail hyn.

Yn olaf, fy nhrydedd rhagdybiaeth yw y bydd y setiau profi acen benodol yn cael effaith negyddol ar y canlyniadau. Mae hyn oherwydd efallai nad yw'r amrywiadau wedi'u cynnwys yn y data hyfforddi yn y profion hyn (V 2.2 Gweithdrefn). Wrth greu brawddegau newydd ar gyfer *Common Voice-cy* ar gyfer y traethawd hir hwn (VI 2.1 Data), roeddwn yn ceisio osgoi cynnwys geiriau tafodieithol er mwyn gallu cynnal prawf teg gyda data Paldaruo ym mhennod VI Math o Ddata. Oherwydd hyn mae'n bosib iawn nad oes amrywiadau yn ymddangos yn y data hyfforddi yn gyfartal. Mae cyflwyno mwy o amrywiadau yn y setiau profi o un acen benodol yn her sy'n debygol o waethygu'r canlyniadau.

2. Dull

Yn y rhan hon, ceir esboniad o'r dull i hyfforddi a phrofi system adnabod lleferydd gyda data acen benodol. Mae cyflwyniad i'r data, yn ogystal ag esboniad o'r broses gaffael o S4C er mwyn cael mynediad at y rhaglenni teledu acen benodol (cyfeirir ato fel data S4C), a'r broses o baratoi'r data yn 2.1. Oherwydd natur y rhaglenni teledu, mae angen golygu'r recordiadau er mwyn dileu unrhyw seibiant neu gerddoriaeth cefndir. Mae hefyd angen tagio'r trawsgrifiadau gyda gwybodaeth ieithyddol oherwydd anghenion *Kaldi*, y cit offer gaiff ei ddefnyddio yn y profion hyn. Trafodir yr anghenion hyn yn 2.1, lle mae cofnod o'r camau golygu. Mae esboniad o'r camau profi yn 2.2, a'r camau i ddadansoddi'r canlyniadau yn 2.3. Cyflwynir canlyniadau'r profion yn rhan 3 a gwelir y dadansoddiad yn rhan 4.

Er mwyn cymharu'r data acen benodol yn erbyn ei gilydd, mae'r set o ddata deheuol a set o ddata gogleddol yn cael eu defnyddio, un ar ôl y llall, fel set brofi gan ddefnyddio'r asesiad cynhenid (*instrinsic evaluation*) (90% set hyfforddi a 10% set brofi, fel gwelir ym mhennod II 4 Dulliau i Ymdopi gyda Llai o Ddata, Ffigwr 8). Fel nodir uchod, mae'r system yn cael ei hyfforddi ar wahanol ddata – Paldaruo 4.0 yn unig, a *Common Voice-cy* (13/02/19) sy'n gyfuniad o bromptiau Paldaruo 4.0 a brawddegau llawn *Common Voice-cy* – er mwyn darganfod mwy am anghenion yn wyneb data acen benodol. Mae profion ychwanegol ar gyfer ymchwilio i effaith hyfforddi a phrofi ar ddata S4C yn unig hefyd. Cynhelir y profion ychwanegol hyn i ymchwilio i effaith cyflwyno'r amrywiadau yn y data hyfforddi yn ogystal â'r data profi.

2.1. Data

Dan amgylchiadau ymchwil academaidd, mae modd derbyn caniatâd i ddefnyddio archif S4C. Cefais gyfres o gyfarfodydd gyda chyswllt yn archif S4C ym Mehefin 2018 er mwyn trafod defnyddio recordiadau'r archif oedd wedi'u his-deitlo. Ar ôl y trafodaethau hyn, dewiswyd cyfres yr un o raglenni *Gwaith Cartref* (GC) (2012) a *Tipyn o Stad* (ToS) (2004). Cafodd y rhain eu dewis gan eu bod, yn gyffredinol, yn adlewyrchu acenion de a gogledd, ac felly'n addas er mwyn astudio effaith acen ar y system. Gwelir manylion y data isod yn Tabl 18:

Rhaglen	**Cyfanswm Oriau**	**Cyfartaledd fesul pennod (awr)**	**Dyddiadau darlledu**
Gwaith Cartref	08:51:21	00:53:08	18/03/12-20/05/12
Tipyn o Stad	07:13:15	00:27:05	17/04/04-31/07/04
Cyfanswm *cyn golygu*	16:04:36	/	/

Tabl 18 Manylion data S4C

Mae GC wedi'i leoli yng Nghaerdydd (de-ddwyrain) mewn ysgol ac yn dilyn bywyd ysgol a chymdeithasol y disgyblion ac athrawon. Mae'r ffeiliau yma yn dod o ail gyfres y rhaglen, penodau 1-10, ac yn cynnwys acenion deheuol gan fwyaf. Mae ToS wedi'i leoli yng Nghaernarfon, (gogledd-orllewin), ac yn dilyn hanes ystâd dai a'r teuluoedd sy'n byw ynddi. Mae'r ffeiliau yn dod o drydedd gyfres y rhaglen, penodau 1 i 16. Clywir acenion gogleddol gan fwyaf ynddi. Fel nodir yn 2 uchod, defnyddir setiau hyfforddi gwahanol ar gyfer y profion cychwynnol. Mae'r rhain yn cynnwys cyfuniadau gwahanol o ddata Paldaruo 4.0, a *Common Voice-cy* (fel gwelir ynTabl 20).

Fel nodir uchod, mae'r deialog yn y rhaglenni yn adlewyrchu acenion ardaloedd Caerdydd (GC) a Chaernarfon (ToS), sy'n golygu bod modd eu defnyddio i fesur effaith data acen ddeheuol a gogleddol ar system adnabod lleferydd. Yn ogystal â'r amrywiaeth sain fel nodir uchod ym mhennod IV 3.3 Tafodiaith ac Acen, mae amrywiaeth o ran geiriau a chystrawen i'w weld hefyd, sy'n amrywio yn y ddwy set. Gwelir rhai enghreifftiau o'r ddwy raglen sy'n dangos amrywio geiriol isod: 'ffaelu' sy'n air tafodieithol am 'methu' yn y de (yn fras), a 'rwan' sy'n air tafodieithol am 'nawr' yn y gogledd (yn fras):

(28) a. 'Wi 'n ffaelu.' (GC) [DE]

b. 'Rhy hwyr rwan yndydy.' (ToS) [GOGLEDD]

Er mwyn cymharu beth yw effaith data o'r de a'r gogledd ar system adnabod lleferydd, defnyddir y data hwn, wedi'i rannu fewn i setiau de a gogledd, a'u profi ar wahân.

Paratoi'r data

Yn yr is-ran hon, ceir esboniad o sut i baratoi data'r rhaglenni teledu S4C er mwyn bod yn addas ar gyfer system adnabod lleferydd. Y camau cyntaf yw:

1. trosi'r ffeiliau testun i fformat .srt (*SubRip*);
2. trosi'r ffeiliau sain i fformat .wav;
3. sicrhau bod y gyfradd ddidol (*bit rate*) yn cyfateb â'r data hyfforddi (er enghraifft, Paldaruo: 16,000 sampl yr eiliad).

Gellir gwneud hyn yn rhaglen VLC. Y cam nesaf yw:

4. creu .textgrid ar gyfer bob pennod.

Nid yw'r broses hon yn creu trawsgrifiad perffaith yn awtomatig yng nghyd-destun bod peiriant adnabod lleferydd yn trawsgrifio. Felly, mae angen golygu'r trawsgrifiadau o fewn rhaglen *Praat* (Boersma & Weenink, 2007). Yn y rhaglen hon, mae modd golygu aliniad y sain a'r testun i sicrhau cywirdeb. Mae hwn yn dasg swmpus felly cyflogwyd intern ym mis Ionawr 2019 i helpu gyda'r gwaith. Mae angen golygu'r testun ym mhob pennod er mwyn gwirio'r alinio rhwng y sain a'r testun. Dylid ychwanegu unrhyw ddeialog ychwanegol hefyd. Dylid 'tagio' unrhyw gerddoriaeth cefndir er mwyn ei ddiystyru o'r profion, yn ogystal â nodi unrhyw acen annisgwyl, hynny yw acen ddeheuol yn ToS neu acen ogleddol yn GC. Er mwyn defnyddio'r cit offer *Kaldi*, dylid 'tagio' rhyw'r siaradwr hefyd gan bod angen y metadata yma ar *Kaldi*. Ar ôl gorffen golygu, mae angen defnyddio sgriptiau pellach i drosi'r ffeiliau .textgrid (wedi'u golygu) i fod yn addas ar gyfer profi o fewn *Kaldi* (a *DeepSpeech*). Hynny yw, rhannu'r ffeiliau .wav a .textgrid i nifer o ffeiliau .wav a .textgrid llai ar gyfer bob segment, a chreu set profi newydd sy'n gweithio o fewn *Kaldi* (a *DeepSpeech*).

Nodir uchod yn 1.1, bod geiriadur ynganu yn cyfrannu at y broses o adnabod lleferydd. Pan nad yw'r system yn gallu adnabod gair, caiff ei nodi mewn ffeil OOV (allan o'r eirfa). Rhai enghreifftiau o gynnwys yr OOV ar gyfer data S4C yw "1", "£" a

“A470”. Problem symbolau yw hyn felly mae angen ychwanegu’r rhain at eiriadur ynganu â llaw fel “un”, “punt” ac “a pedwar saith deg”.

Ar ôl golygu’r data dilys, diystyrir 1771 segment o’r data (4.91 awr GC a 5.29 awr ToS) oherwydd eu bod yn cynnwys cerddoriaeth gefndir, enghraifft o dagiau er enghraifft, ‘<#B>’ am ‘benywaidd’ neu, ddim yn eiriau o gwbl er enghraifft, ‘-’. Roedd cerddoriaeth gefndir yn rhan fawr o hyn oherwydd nifer y golygfeydd lle dim ond cerddoriaeth oedd i’w glywed. Dyma felly yw cyfanswm yr oriau dilys o ddata S4C:

Data	**Oriau *dilys* (awr)**	**Cyfanswm *gwreiddiol* (awr)**	**Colled o’r gwreiddiol (%)**
Gwaith Cartref	2.40	7.31	-67.17%
Tipyn o Stad	3.22	8.51	-62.16%
Cyfanswm	6.02	16.04	-62.46%

Tabl 19 Oriau dilys data S4C

2.2. Gweithdrefn: Profi

Crëir setiau profi hafal (2.40 awr) o’r data hwn er mwyn cynnal prawf teg wrth gymharu effaith y setiau acen benodol yn erbyn ei gilydd. I baratoi’r data fel setiau profi ar gyfer y profion hyn, dilynir camau sy’n cael eu disgrifio ym mhenodau V 2.2 Gweithdrefn a VI 2.2 Gweithdrefn, i greu ffeiliau .csv. Defnyddir data Paldaruo 4.0, a *Common Voice-cy* (13/02/19), a chyfuniad o’r ddau, fel setiau hyfforddi sydd wedi’u paratoi ar gyfer y cit offer *Kaldi* yn barod. Defnyddir yr asesiad cynhenid (*instrinsic evaluation*) (II 4 Dulliau i Ymdopi gyda Llai o Ddata, Ffigwr 8) er mwyn profi data Paldaruo 4.0 a *Common Voice-cy* (13/02/19) ar ddata acen benodol S4C. Cynhelir 18 prawf er mwyn profi effaith y setiau profi gogledd, de a chyfuniad o’r ddau ar y data canlynol, fel gwelir yn Tabl 20:

1. effaith hyfforddi ar **eiriau unigol** (Paldaruo 4.0);
2. effaith hyfforddi ar **frawddegau llawn a geiriau unigol** (*Common Voice-cy* (13/02/19));
3. effaith hyfforddi ar ddata **mwy** mewn maint (*Common Voice-cy* (sy'n cynnwys promptiau Paldaruo 4.0) + data Paldaruo 4.0);
4. effaith hyfforddi ar ddata **acen ddeheuol** (De);
5. effaith hyfforddi ar ddata **acen ogleddol** (Gogledd);
6. effaith hyfforddi ar y **cyfuniad** o'r data acen benodol (De+Gogledd).

	Kaldi	*Oriau*	**Set Brofi**		
Set Hyfforddi	Paldaruo 4.0	38	De	Gogledd	De + Gogledd
	Common Voice-cy (13/02/19)	42	De	Gogledd	De + Gogledd
	Common Voice-cy (13/02/19) +Paldaruo 4.0	80	De	Gogledd	De + Gogledd
	De	2.40	De	Gogledd	De + Gogledd
	Gogledd	2.40	De	Gogledd	De + Gogledd
	De + Gogledd	5.20	De	Gogledd	De + Gogledd

Tabl 20 Profion i'w cynnal gyda data S4C

Ar gyfer y prawf olaf (set hyfforddi: De+Gogledd; Set brofi: De+Gogledd), defnyddir y dull k-*fold cross-validation* (C-V) (II 4 Dulliau i Ymdopi gyda Llai o Ddata, Ffigwr 9) i rannu data S4C yn set hyfforddi a phrofi ar hap. Defnyddir yr asesiad cynhenid (*instrinsic evaluation*) (rhannu'r corpws cyfan yn set hyfforddi (90%) a set brofi (10%), II 4 Dulliau i Ymdopi gyda Llai o Ddata, Ffigwr 8) er mwyn rhedeg y prawf er mwyn cyfateb i weddill y profion yn y bennod hon.

2.3. Gweithdrefn: Dadansoddi

Er mwyn dadansoddi'r canlyniadau, dilynir camau Meyer (2019) er mwyn asesu perfformiad modelau *Kaldi*. Gwneir hyn drwy edrych ar y ffeil log 'best_path' a dadansoddi'r set gyfan ar gyfer y de a'r gogledd. Mae'r ffeil log yma'n cofnodi 'cost' bob segment neu frawddeg. Yn ôl dogfennaeth *Kaldi* (Povey, et al., 2011), mae 'best path' yn golygu'r gost isaf posib er mwyn cyrraedd yr ateb mwyaf tebygol, felly yr isaf yw'r gost, y gorau oll fydd y canlyniad. Er mwyn mesur a chymharu'r cofnodion hyn, defnyddir y *Pellter Levenshtein* (*Levenshtein Distance*) neu'r *Pellter Golygu* (*Edit Distance*). Mae

Levenshtein yn mesur faint o fewnosodiadau, dileadau ac amnewidiadau sydd angen ar gyfer creu'r allbwn o'r mewnbwn (Levenshtein, 1966). Er enghraifft, y pellter *Levenshtein* rhwng 'mêl' a 'sêr' yw 2:

1. mêl > sêl (mewnosodiad 's')
2. sêl > sêr (mewnosodiad 'r')

Oherwydd bod *Levenshtein* yn mesur mewn llythrennau er mwyn cyfri'r 'camau' (llythrennau) sydd angen newid i gyrraedd yr allbwn, defnyddir cyfradd llythrennau gwallus (*character error rate*, *CER*) i ddadansoddi'r canlyniadau hyn. Mae *Levenshtein* yn algorithm poblogaidd yn y maes prosesu signal, gan gynnwys adnabod lleferydd.

Er mwyn deall mwy am beth sy'n achosi problem yn y data, y cam nesaf yw dadansoddi'r allbwn yn thematig. Drwy ddefnyddio dadansoddiad thematig, mae modd categoreiddio mathau gwahanol o wallau a gweld os oes patrwm yn amlygu. Defnyddiwyd safonau ISO (2018) er mwyn penderfynu ar y categorïau a'i diffinio, a *Geiriadur Prifysgol Cymru* i adnabod geiriau fel cyfnewid cod; gwelir isod.

1. **Cywair**
   Ffurfiol: cydymffurfio â safonau safonol, defnyddir mewn sefyllfaoedd difrifol a swyddogol;
   anffurfiol: ddim yn cydymffurfio â safonau safonol, defnyddir mewn sefyllfaoedd achlysurol ac answyddogol;
   sathredig: cywair hynod anffurfiol megis rhegi, defnyddir ar lafar mewn iaith bob dydd, llai cyffredin mewn dogfennaeth.
2. **Enwau endidau** (*named entities*)
   Mr, Mrs, Tesco, ac yn y blaen.
3. **Cyfnewid cod**
   Cyfnewid rhwng 2 neu fwy o ieithoedd o fewn sgwrs.
4. **Tafodiaith**
   Amrywiadau mewn geirfa neu gystrawen yn ôl ardal.

Defnyddiwyd Macros Excel i godio'r gwahanol gategorïau hyn a gweld pa mor aml mae pob un yn achosi gwall, yn ogystal ag ansawdd yr allbwn dan sylw.

3. Canlyniadau

Yn y rhan hon, ceir canlyniadau WER y profion yn gyntaf yn 3.1, yna dadansoddiad pellach o'r gwallau wrth adnabod lleferydd data S4C yn 3.2.

3.1. Canlyniadau WER

Er mwyn darganfod beth yw effaith y newidyn acen ar system adnabod lleferydd Cymraeg, cynhelir yr 18 prawf yma gan ddefnyddio cyfuniadau o ddata gwahanol fel gwelir yn Tabl 20 (2.2). Gwelir canlyniadau'r profion yn Tabl 21 isod. Nid oes cofnod o wyriad safonol ac yn y blaen fel sydd ym mhennod V Citiau Offer. Mae hyn oherwydd mai'r asesiad cynhenid (*instrinsic evaluation*) sy'n cael ei ddefnyddio i gyfrifo mwyafrif y canlyniadau yn y bennod hon, yn hytrach na'r dull C-V (k-*fold cross-validation*) lle mae angen gwirio canlyniadau bob plyg drwy wiriad safonol ac yn y blaen.

WER% (*Kaldi*)			**Set Brofi**		
		Oriau	S4C De	S4C Gogledd	S4C De & Gogledd
Set Hyfforddi	Paldaruo 4.0	38	98.93	99.73	99.66
	Common Voice-cy (13/02/19)	42	85.31	87.60	85.84
	Common Voice-cy (13/02/19) + Paldaruo 4.0	80	90.46	90.86	90.77
	S4C De	2.40	65.04	76.24	75.91
	S4C Gogledd	2.40	77.89	65.49	77.81
	S4C De & Gogledd	5.20	60.15	60.94	59.92

Tabl 21 Profi data acen benodol S4C ar setiau hyfforddi gwahanol Paldaruo 4.0, *Common Voice-cy* (13/02/19), *Common Voice-cy* (13/02/19)+Paldaruo 4.0, ac ar ddata S4C (de, gogledd, de+gogledd)

Mae'r canlyniadau yn Tabl 21 yn edrych ar WER y profion gwahanol wrth gael eu profi ar ddata acen benodol S4C. Wrth edrych ar y tabl yn llorweddol, mae *Common Voice-cy* yn rhoi canlyniadau gwell o 13.19% WER (ar gyfartaledd), wrth gymharu gyda chanlyniadau Paldaruo 4.0 yn ymdopi gydag amrywio acen. Mae hyn yn dangos un o ddau beth: gan fod 4 awr ychwanegol gan *Common Voice-cy* wrth gymharu gyda 38 awr Paldaruo 4.0, mae'n bosib bod y system yn ymateb yn well i *faint* mwy o ddata; neu,

mae'n bosib mai'r *math* o ddata (VI Math o Ddata) sydd yn *Common Voice-cy* – brawddegau llawn – sy'n cael yr effaith fwyaf. Mae canlyniadau'r set hyfforddi *Common Voice-cy*+Paldaruo 4.0, yn awgrymu bod y *math* o bromptiau sy'n cael ei ddefnyddio yn cael effaith mwy ar y canlyniadau. Gwelir hyn gan fod hyfforddi ar ddata *Common Voice-cy*+Paldaruo 4.0, sy'n gymysgedd o bromptiau geiriau unigol Paldaruo 4.0 a brawddegau llawn, gydag ychwanegiad o ddata geiriau unigol o Paldaruo 4.0, yn gwaethygu'r canlyniadau o 4.45% WER ar gyfartaledd. Daw'r ffigwr 4.45% WER wrth gymryd cyfartaledd y gwahaniaeth rhwng holl ganlyniadau hyfforddi ar gyfuniad Paldaruo 4.0+*Common Voice-cy*, a holl ganlyniadau hyfforddi ar ddata *Common Voice-cy* yn unig. Ceir y gwaethygiad hwn er gwaethaf maint sylweddol fwy y set *Common Voice-cy*+Paldaruo 4.0 (80 awr). Dylid nodi bod y gwahaniaeth rhwng canlyniadau *Common Voice-cy*+Paldaruo 4.0 a *Common Voice-cy* yn unig gyn lleied â 5.15% WER ar ei fwyaf, felly nid oes gwahaniaeth mawr wrth ychwanegu data Paldaruo 4.0 eto i *Common Voice-cy* fel set hyfforddi.

Wrth edrych ar Tabl 21 yn fertigol, does dim llawer iawn o wahaniaeth yn y WER wrth brofi ar y de neu'r gogledd (acen ogleddol yn +1.16% WER ar gyfartaledd o gymharu gyda chanlyniadau profi ar acen ddeheuol), wedi'i hyfforddi ar ddata Paldaruo 4.0, *Common Voice-cy*, neu *Common Voice-cy*+Paldaruo 4.0. I atgoffa, wrth hyfforddi a phrofi ar ddata *Common Voice-cy* yn unig, VI 3 Canlyniadau, Tabl 16, roedd y WER yn 3.44% felly mae'r naid sydd i'w weld yng nghanlyniadau WER Tabl 21 yn uchel iawn. Mae'n amlwg felly bod rhywbeth yn y setiau profi S4C yn achosi gwallau. Gan nad oes gwahaniaeth mawr *rhwng* yr acenion, mae'n awgrymu bod elfennau eraill yn y data S4C yn broblemus. Cynhelir dadansoddiad pellach yn 3.2 i ddeall mwy am ba elfennau sy'n cyfrannu at WER mor uchel wrth brofi ar ddata S4C.

3.2. Dadansoddi Gwallau

Yn yr is-ran hon, edrychir ar nifer y gwallau sy'n cyfrannu at WER uchel canlyniadau Tabl 21. Gwelir dadansoddiad o'r gwallau sy'n codi yng nghanlyniadau *Common Voice-cy* isod yn Tabl 22. Daw'r canlyniadau hyn o brawf ar ddata S4C de a gogledd sydd wedi'i hyfforddi ar ddata *Common Voice-cy.* Defnyddir y prawf hwn fel esiampl gan mai hwn yw'r canlyniad gorau wrth brofi'r system ar ddata acen benodol, heb gael ei hyfforddi arno (Tabl 21). Fel nodir yn 2.3, defnyddir y ffeil log 'best_path' er mwyn edrych yn fanylach ar y canlyniadau.

% y gwallau yn yr allbwn	0-24%	25-49%	50-74%	>75%
De	81	73	113	238
% o gyfanswm y set ddeheuol	16.03%	14.46%	22.38%	47.13%
Gogledd	65	67	147	226
% o gyfanswm y set ogleddol	12.87%	13.27%	29.11%	44.75%

Tabl 22 Graddfa gwallau wrth adnabod data S4C (hyfforddi ar ddata *Common Voice-cy*)

Mae cyfanswm o 505 allbwn o'r set ddeheuol a 505 allbwn o'r set ogleddol yn y ffeil log hwn. Wrth ddefnyddio'r *Pellter Levenshtein*, mae modd gweld yr allbynnau hyn ar raddfa cywirdeb lle mae 0% yn dangos dim gwall o gwbl yn yr allbwn a 100% yn dangos pob gair yn yr allbwn wedi'i gam-adnabod. Wrth edrych ar yr allbynnau cywirdeb gwaethaf, >75%, does dim gwahaniaeth mawr rhwng y de a'r gogledd (de=238, gogledd=226), ond nodir mai dyma lle mae'r nifer fwyaf o wallau yn y set gyfan. Mae hyn yn meddwl bod 47.13% o'r allbynnau gwallus de, a 44.75% o'r allbynnau gwallus gogledd, yn cynnwys dros 75% o wallau yn yr allbwn. Wrth edrych ar yr allbynnau cywirdeb 50-74% (de=113, gogledd=147), yr allbynnau cywirdeb 25-49% (de=73, gogledd 67), a'r allbynnau cywirdeb 0-24% (de=81, gogledd=65), does dim gwahaniaeth mawr i'w weld rhwng y de a'r gogledd chwaith. Fodd bynnag, mae'r set ogleddol i'w weld yn achosi mwy o wallau yn yr allbynnau mwyaf gwallus (50-74%, >75%), sy'n awgrymu bod y system yn ymdopi ychydig yn well gydag acen y de. Mae hyn yn cefnogi canlyniadau Tabl 21, bod setiau acen y gogledd yn peri ychydig mwy o drafferth na setiau acen y de wrth brosesu. Fel nodir yn 3.1, gan nad oes llawer iawn o wahaniaeth rhwng y canlyniadau acen de a gogledd, edrychir yma i elfennau eraill all fod yn cyfrannu at yr WER uwch. Yn y rhan nesaf, ceir ystyriaeth o'r elfennau eraill hyn all fod yn gwaethygu'r canlyniadau hyn.

Amrywio Iaith

Gan fod canlyniadau Tabl 21 yn dangos nad oes gwahaniaeth mawr rhwng acenion, edrychir ar 4 categori ychwanegol yn y dadansoddiad isod. Yma, arbrofir gyda'r 4 categori er mwyn deall mwy am beth sy'n achosi'r problemau o fewn setiau de a gogledd S4C. I atgoffa, dewiswyd y categorïau hyn o safonau ISO (2018) a defnyddir y diffiniadau o'r safonau hynny. Gwelir y categorïau isod gydag enghreifftiau cyfatebol:

1. **Cywair**

   Ffurfiol: cydymffurfio â safonau safonol, defnyddir mewn sefyllfaoedd difrifol a swyddogol;

   anffurfiol: ddim yn cydymffurfio â safonau safonol, defnyddir mewn sefyllfaoedd achlysurol ac answyddogol;

   sathredig: cywair hynod anffurfiol, defnyddir ar lafar mewn iaith bob dydd, llai cyffredin mewn dogfennaeth. Er enghraifft:

   MAE HYNNA 'N BETH DA NAGYW E (GC) [DE, CYWAIR ANFFURFIOL]

   BASTARD ‘SDIM RHAID I FI GADW’R BABI ‘MA STI (ToS) [GOGLEDD, CYWAIR SATHREDIG]

2. **Enwau endidau**

   Mr, Mrs, Tesco, ac yn y blaen, er enghraifft:

   FALLE DYLAI MR WATKINS GYMRYD DROSTO (GC*)* [DE, ENWAU ENDIDAU]

   DACH CHI 'SHO GADA'L NEGAS MRS REES (ToS) [GOGLEDD, ENWAU ENDIDAU]

3. **Cyfnewid cod**

   Cyfnewid rhwng 2 neu fwy o ieithoedd o fewn sgwrs, er enghraifft:

   DW I ANGEN FY BEAUTY SLEEP (GC) [DE, CYFNEWID COD]

   REIT PASIA’R SCREWDRIVER I MI ‘NEI DI (ToS) [GOGLEDD, CYFNEWID COD]

4. **Tafodiaith**

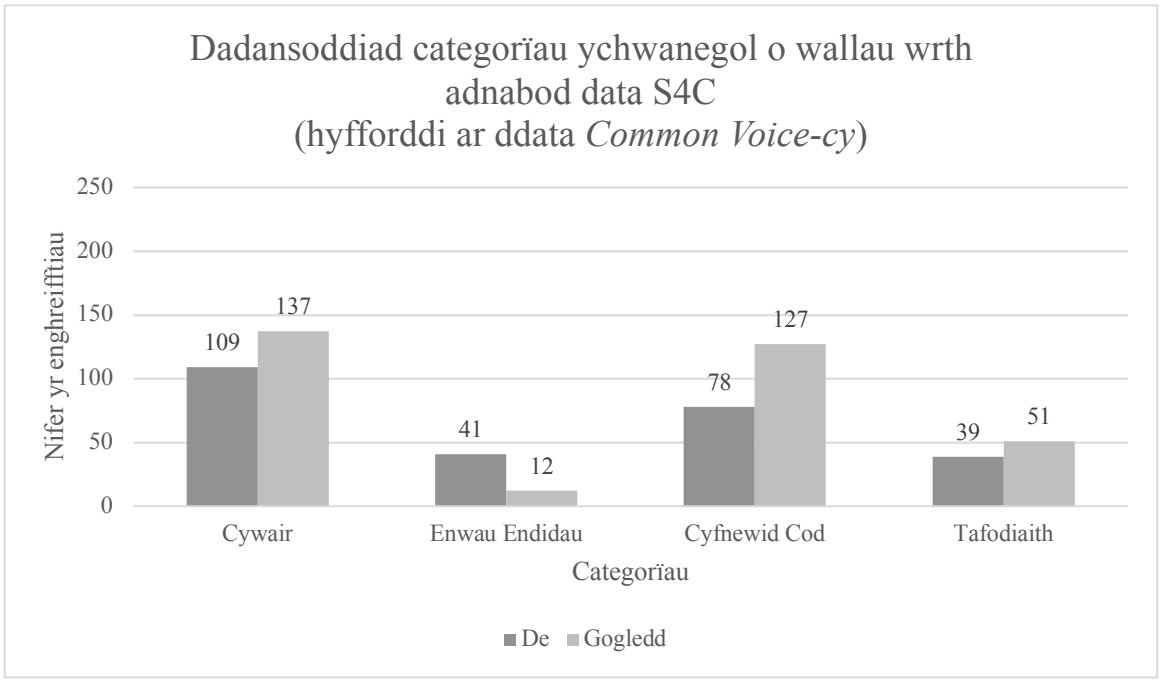
   Amrywiadau mewn geirfa neu gystrawen yn ôl ardal, er enghraifft:

   'SDIM RHESWM 'DA TI I FECSO (GC) [DE, TAFODIAITH]

   SA R'WBATH BLAW RHAI CAWS RHYFADD 'MA (ToS) [GOGLEDD, TAFODIAITH]

Yn y dadansoddiad isod, nid yw cyfanswm y gwallau yn hafal i gyfanswm y setiau oherwydd nad oes rheswm amlwg dros bob gwall (dim ond gwallau wedi’u

categoreiddio sydd wedi'u cynnwys yma). Mae modd categoreiddio 267 (52.9%) gwall yn set y de a 327 (64.8%) gwall yn set y gogledd yn ôl y categorïau uchod.



Ffigwr 19 Dadansoddiad categorïau ychwanegol o wallau wrth adnabod data S4C (hyfforddi ar ddata *Common Voice-cy*)

Mae Ffigwr 19 yn dangos mai cywair anffurfiol (41.41% o gyfanswm y gwallau deheuol a gogleddol wedi'u categoreiddio uchod, 246 o 594 gwall dros y ddwy set) a chyfnewid cod (34.51% o'r gwallau deheuol a gogleddol wedi'u categoreiddio uchod, 205 o 594 gwall dros y ddwy set) sy'n achosi'r rhan fwyaf o'r gwallau yn y categorïau hyn. Mae'r rhain yn cyfeirio at amrywio mewn defnydd o iaith (cywair anffurfiol) neu ddefnydd o iaith Saesneg mewn sgwrs Gymraeg (cyfnewid cod). Dengys Ffigwr 19 bod enwau endidau dim ond yn cyfrif am 8.92% o'r gwallau uchod (8.92% o gyfanswm y gwallau deheuol a gogleddol wedi'u categoreiddio uchod, 51 o 594 gwall dros y ddwy set) a thafodiaith yn cyfrif am 15.15% o'r gwallau uchod (15.15% o gyfanswm y gwallau deheuol a gogleddol wedi'u categoreiddio uchod, 90 o 594 gwall dros y ddwy set).

Mae'r dadansoddiad yn Ffigwr 19 yn dangos bod y categorïau uchod yn cyfrannu at y gwallau wrth brosesu data S4C. Mae hyn yn golygu nad dim ond amrywiadau acen sy'n cyfrannu at WER uchel a welir yn Tabl 21. Dengys hyn fod y darlun yn fwy cymhleth,

gydag amrywiaeth o nodweddion megis cywair anffurfiol a chyfnewid cod yn effeithio'r canlyniadau.

4. Trafodaeth

Yn y rhan hon, ceir trafodaeth o ganlyniadau 3 gan gynnwys canlyniadau WER 3.1 a dadansoddiad gwallau 3.2 uchod. Wrth edrych ar ganlyniadau Tabl 21 uchod, awgrymir bod data *Common Voice-cy* (13/02/19) yn ymdopi'n well na data Paldaruo 4.0 wrth gyflwyno acen i'r set brofi. Dengys hyn mai'r math o ddata sy'n cael yr effaith fwyaf, gan fod brawddegau llawn a geiriau unigol yn *Common Voice-cy* (42 awr) yn hytrach na geiriau unigol yn unig yn Paldaruo 4.0 (38 awr). Er, ychydig iawn o wahaniaeth sydd rhwng canlyniadau data *Common Voice-cy* wedi'i gyfuno efo data Paldaruo 4.0 eto (80 awr) (4.45% WER ar gyfartaledd). Wrth gymharu effaith acen y de a'r gogledd ar ddata *Common Voice-cy*, Paldauro 4.0 a'r cyfuniad o hyn gydag ychwanegiad Paldaruo 4.0 eto, does dim gwahaniaeth mawr rhwng y canlyniadau. Gan fod yr WER mor uchel o gymharu gyda chanlyniadau hyfforddi a phrofi ar setiau *Common Voice-cy* yn unig (pennod VI 3 Canlyniadau, Tabl 16), mae'n amlwg bod rhywbeth am setiau S4C yn achosi problem. Oherwydd nad oes gwahaniaeth mawr rhwng y canlyniadau acen de a gogledd, ymddengys nad acen yw'r brif broblem, ond cymysgedd o amrywiadau gwahanol. Felly, dadansoddir y canlyniadau ymhellach yn 3.2 er mwyn deall mwy am rhai o'r gwallau yn y setiau de a gogledd. Dengys dadansoddiad Ffigwr 19 uchod bod cymysgedd o gywair anffurfiol a chyfnewid cod yn achosi nifer fawr o'r gwallau wrth adnabod data S4C. Yn y rhan hon, ceir trafodaeth o'r amrywiadau sydd i'w canfod yn y setiau profi S4C, gan gynnwys goblygiadau'r canfyddiadau hyn.

Fel nodir yn 3.2, edrychir i 4 categori pellach: cywair anffurfiol, cyfnewid cod, enwau endidau a, thafodiaith. Mae natur sgyrsiol rhaglenni GC a ToS yn golygu bod mwy o gywair anffurfiol yn ymddangos yn y deialog rhwng y cymeriadau. Noda Thomas & Webb-Davies (2017, t. 48) bod cyfnewid cod yn 'nodwedd ar iaith lafar anffurfiol' felly mae'n dilyn fod llawer o enghreifftiau o gyfnewid cod yn y rhaglenni hefyd. Mae'n debygol bod mwy o enghreifftiau o gywair anffurfiol (ac felly cyfnewid cod o bosib) yn y set ogleddol oherwydd lleoliad y rhaglen, sef ystâd dai. Ar y llaw arall, mae llawer mwy o enwau endidau i'w clywed yn set y de, oherwydd lleoliad y rhaglen mewn ysgol. Gan fod GC, sef y data deheuol, wedi'i leoli mewn ysgol, mae cyfeiriadau at yr athrawon yn aml ar ffurf Mr, neu Mrs er enghraifft, sy'n cael eu cyfri fel enwau endidau. Gan nad oes llawer o sefyllfaoedd ffurfiol yn ToS, does dim cymaint o enwau endidau fel hyn yn ymddangos. Does dim gwahaniaeth mawr rhwng nifer y gwallau tafodieithol yn y set ddeheuol nac yn y set ogleddol.

Gan fod yr amrywio iaith yn y categorïau yn amrywio geirffurfiau'r iaith yn ogystal â'r sain, mae angen ystyried pa fath o ddata sy'n cael ei gyflwyno fel set brofi, a beth yw rôl y model iaith yn ogystal â'r model acwstig (1.1). Mae angen edrych i rôl y model iaith yma yn sgil hyn gan fod dadansoddiad Ffigwr 19 yn dangos bod amrywiadau cywair a chyfnewid cod yn cyfrannu at nifer y gwallau, yn ogystal ag acen, ar ganlyniadau Paldaruo 4.0 a *Common Voice-cy*. Mae'r model iaith yn cael ei greu o'r data hyfforddi yn unig yn y profion hyn. Felly, os caiff rhai amrywiadau iaith eu cyflwyno yn y data profi yn unig, nid ydynt wedi'u cynnwys yn y model iaith, felly nid yw'r system yn debygol o'u hadnabod yn llwyddiannus. Mae'n debygol bod hyn yn cyfrannu at y gwallau uchod gan nad yw'r enghreifftiau o gywair anffurfiol neu gyfnewid cod yn ymddangos yn y data *Common Voice-cy.* Mae gwaith ar ieithoedd eraill wedi ymchwilio i effaith cyflwyno agweddau newydd fel hyn i'r model iaith. Gweler gwaith Yılmaz et al. (2016) hefyd ar gyflwyno cyfnewid cod gyda model iaith ddwyieithog Frisieg-Iseldireg (gwelliant o 14.45%, o 69.2% i 59.5% WER). Mae hyn yn dangos bod angen gwaith pellach ar arbrofi gyda dulliau gwahanol o greu model iaith mwy cadarn, yn ogystal â gwella'r model acwstig gan ychwanegu mwy o ddata. Byddai hyn yn edrych i weld os byddai gwelliant yn y canlyniadau wrth adnabod amrywiadau iaith. Caiff hyn ei drafod ymhellach ym mhennod IX Trafodaeth.

I grynhoi, dengys Tabl 21 nad yw canlyniadau y profion sy'n hyfforddi gyda data Paldaruo 4.0, *Common Voice-cy*, neu *Common Voice-cy*+Paldaruo 4.0, yn cefnogi fy rhagdybiaeth gyntaf. Dyma'r rhagdybiaeth sy'n rhagdybio y byddai acen y gogledd yn cael effaith gwaeth nac acen y de. Nid yw'r canlyniadau'n cefnogi'r rhagdybiaeth gan nad oes gwahaniaeth mawr i'w weld rhwng acenion. Mae'r gwahaniaeth mwyaf dim ond yn dangos gwahaniaeth o 2.29% WER (de 85.31% WER, gogledd 87.60% WER wrth hyfforddi ar ddata *Common Voice-cy*). Er bod hyn yn ychydig iawn o wahaniaeth, does bron i ddim gwahaniaeth wrth hyfforddi ar Paldaruo 4.0 (38 awr) neu gyda *Common Voice-cy*+Paldaruo 4.0 (80 awr). Nid yw hyn yn cefnogi fy ail ragdybiaeth y byddai Paldaruo 4.0 yn ymdopi'n well gyda'r acen ogleddol oherwydd nifer y cyfranwyr gydag acen ogleddol. Mae'r canlyniadau'n awgrymu bod data acen ogleddol ychydig yn anoddach i brosesu gan ei fod yn rhoi WER ychydig yn uwch (1.16% WER ar gyfartaledd o gymharu gyda chanlyniadau profi ar acen ddeheuol Paldaruo 4.0, *Common Voice-cy*, *Common Voice-cy*+Paldaruo 4.0), ond nid yw'r gwahaniaeth yn fawr. Mae dadansoddiad Ffigwr 19 yn dangos bod rhesymau eraill am yr anhawster prosesu (cywair anffurfiol, cyfnewid cod ac yn y blaen). Wrth hyfforddi ar ddata S4C yn unig, mae'r

canlyniadau yn cefnogi fy nhrydedd ragdybiaeth hefyd. Mae hon yn rhagdybio bod cyflwyno amrywiadau newydd yn y data profi yn unig yn cael effaith negyddol ar y canlyniadau. Wrth hyfforddi ar ddata S4C yn unig yn y profion hyn, cyflwynir mwy o'r amrywiadau yn y data hyfforddi, a gwelir gwelliant o 29.76% (85.31% WER i 59.92% WER) wrth gymharu canlyniad gorau *Common Voice-cy* gyda data S4C (85.31% WER), a chanlyniadau gorau S4C ar ei hun (59.92% WER, de & gogledd hyfforddi a phrofi). Mae hyn yn awgrymu bod cyflwyno amrywiadau iaith yn y data hyfforddi yn gwella gallu system i adnabod lleferydd sgyrsiol fel sydd yn nata S4C. Bydd goblygiadau hyn yn cael eu trafod ymhellach ym mhennod IX Trafodaeth.

VIII. *DeepSpeech*

1. Cyflwyniad

Pwrpas y bennod hon yw amlinellu'r profion a gynhelir gyda pheiriant *DeepSpeech* er mwyn adeiladu ar bennod V Citiau Offer wrth ymchwilio i gwestiwn ymchwil 1, ynghylch pa ffordd o fodelu'r Gymraeg yn gyfrifiadurol sy'n creu'r system adnabod lleferydd mwyaf cadarn. Mae'n edrych i sut mae peiriant *DeepSpeech* yn ymdopi gydag adnabod lleferydd Cymraeg sy'n dioddef o ddiffyg data (I 2.4 Diffyg Data, Amrywio Ieithyddol a'r Dulliau Perthnasol). Pwrpas y profion hyn yn benodol yw darganfod sut mae modelau pen-i-ben *DeepSpeech* yn ymdopi gyda diffyg data fel sydd ar gyfer y Gymraeg. Ceir cyflwyniad i fodelau pen-i-ben ym mhennod II 3.3 Modelau Pen-i-ben. Yn y bennod hon, ceir esboniad o'r profion a gynhelir er mwyn profi ymarferoldeb y rhagdybiaeth berthnasol sydd i'w gweld yn 1.3. Unwaith eto, mae'r bennod wedi'i rhannu i dair rhan: Dull (Data, Gweithdrefn), Canlyniadau a Thrafodaeth. Bydd trafodaeth gyffredinol am oblygiadau ac arwyddocâd y canlyniadau ym mhennod IX Trafodaeth.

1.1. Yr Her: Diffyg Data a'r Dyfodol

Mae'r profion yn y bennod hon yn ffocysu ar ddiffyg data. Mae diffyg data yn gyffredin ymysg ieithoedd llai eu hadnoddau fel y Gymraeg (gwelir pennod I 2.4 Diffyg Data, Amrywio Ieithyddol a'r Dulliau Perthnasol, a phennod V 1.1 Yr Her: Diffyg Data). Cyfeiria hyn at ieithoedd sydd heb ddigonedd o ddata sain addas. Gwelir enghreifftiau o gorpora 'digonol' yn Saesneg, yn Tabl 23, (gydag eraill yn Tabl 24) wedi'u cymharu gyda'r Gymraeg, sydd ond ag 80 awr ar adeg profi (wedi seilio ar ddata Paldaruo 4.0 a *Common Voice-cy* (13/02/19)).

	Corpws	Oriau
Saesneg	*Librispeech*	tua 1,000
	Fisher	tua 2,000
Cymraeg	Paldaruo 4.0	38
	Common Voice-cy (13/02/19)	42

Tabl 23 Cymharu nifer oriau corpora agored Saesneg a Chymraeg

Fel nodir ym mhennod V 1.1 Yr Her: Diffyg Data hefyd, mae'r term 'llai eu hadnoddau' yn gallu cyfeirio at ddiffyg data oherwydd cost a hyd y broses o gasglu data neu gyfyngiadau trwydded. Ond, mae'r term hefyd yn gallu cyfeirio at ddiffyg adnoddau cyllid neu arbenigedd, yn enwedig wrth ddatblygu gwybodaeth ieithyddol. I atgoffa, rhan o hwn yw'r geiriadur ynganu, sy'n gweithredu fel pwynt cyd-destun yn y system brosesu, wedi'i leoli yn y broses ddadgodio, rhwng y mewnbwn a'r allbwn (V 2.2 Gweithdrefn, Ffigwr 16).

Wrth ystyried dyfodol yr her diffyg data ymysg ieithoedd llai eu hadnoddau, mae Mozilla yn cynnig opsiwn newydd. Fel nodir ym mhennod VI 2.1 Data, cyhoeddwyd ymgyrch casglu data cod agored gan Mozilla yn 2017, a gafodd ei ymestyn i gefnogi ieithoedd llai eu hadnoddau, gan gynnwys y Gymraeg, yn 2018 (Ardila, et al., 2020). Mae'r ymgyrch casglu data hon, *Common Voice*, wrthi'n casglu data ar gyfer 38 iaith erbyn 2020 (Ardila, et al., 2020, t. 4218), gan gynnwys ieithoedd llai eu hadnoddau. Mae'n cynnig platfform casglu data cod agored sy'n adnodd allweddol yn y maes ar gyfer ieithoedd llai eu hadnoddau.

Wrth ystyried dyfodol yr heriau eraill sy'n gysylltiedig ag ieithoedd llai eu hadnoddau, megis argaeledd cyllid neu arbenigedd er mwyn datblygu gwybodaeth ieithyddol megis geiriadur ynganu, mae Mozilla yn cynnig opsiwn newydd hefyd. Un o ddatblygiadau mwyaf diweddar y maes adnabod lleferydd yw pensaernïaeth pen-i-ben, sy'n ddull o adeiladu modelau sydd dim ond yn defnyddio'r data. Mae'n gwneud i ffwrdd â gwybodaeth ieithyddol, ac felly'r angen am gomisiynu gwaith arbenigol ieithyddol, fel creu geiriadur ynganu. Mae Mozilla yn defnyddio data *Common Voice* i fwydo mewn i'w beiriant pen-i-ben, *DeepSpeech* (Ardila, et al., 2020, tt. 4220-4221). Mae hwn yn

arddangos y technegau mwyaf diweddar ym maes adnabod lleferydd, ond yn gofyn am ddata mawr.

Mae'r cyfuniad o ymgyrch casglu data cod agored *Common Voice* a phensaernïaeth newydd *DeepSpeech*, sy'n gwneud i ffwrdd â gwybodaeth ieithyddol fel geiriadur ynganu, yn codi'r cwestiwn o sut mae pensaernïaeth pen-i-ben yn ymdopi mewn sefyllfa lai ei hadnoddau megis diffyg data, fel y Gymraeg. Mae timau ymchwil megis Agarwal a Zesch (2020) (Almaeneg y Swistir) a Hjortnæs et al. (2020) (Komi) wedi cynnal arbrofion gydag ieithoedd llai eu hadnoddau eraill lle daethpwyd i'r canlyniad nad oedd *DeepSpeech* yn addas ar gyfer eu hieithoedd nhw. Pwrpas y profion yn y bennod hon yw darganfod sut mae *DeepSpeech* yn ymdopi gyda'r holl ddata Cymraeg sydd wedi cael ei gyflwyno yn y traethawd hir hwn, ac os yw'r dull yn ymarferol ar gyfer y data Cymraeg.

1.2. Cymharu *Kaldi* a *DeepSpeech*

Yn y rhan hon ceir cyflwyniad i *DeepSpeech* yn y cyd-destun o gael ei gymharu gyda chit offer *Kaldi*. Ceir cyflwyniad i adeiladwaith *Kaldi* ym mhennod V 1.2 Cymharu HTK a *Kaldi* a V 2.2 Gweithdrefn, lle mae esboniad o'i bensaernïaeth groesryw. I atgoffa, mae *Kaldi* yn defnyddio pensaernïaeth groesryw HMM-DNN i adeiladu modelau wrth ddefnyddio gwybodaeth ieithyddol megis geiriadur ynganu. Mae'r wybodaeth ieithyddol hwn yn helpu dadgodio'r sain ond, fel nodir uchod, gall fod yn broses hir a drud sy'n afrealistig ar gyfer ieithoedd llai eu hadnoddau. Ar gyfer profion *Kaldi* y traethawd hir hwn, defnyddir geiriadur ynganu o'r project GALLU a gwelir manylion hwnnw ym mhennod III 2.3 GALLU, Paldaruo a Macsen.

Hanes DeepSpeech a Deep Speech

Amrywiaeth Mozilla o *DeepSpeech* yw'r gweithrediad cod agored cyntaf o bapur labordy ymchwil Baidu ar *Deep Speech* (Hannun, et al., 2014). Baidu oedd y cyntaf i gyflwyno'r cysyniad o ddefnyddio rhwydweithiau niwral (*neural networks*) mewn peiriant pen-i-ben, dan y teitl *Deep Speech* (Hannun, et al., 2014). Dechreuodd waith y labordy yn 2014, pan gyhoeddwyd cysyniad newydd o ddefnyddio rhwydweithiau niwral artiffisial o fewn peiriant pen-i-ben er mwyn adnabod lleferydd. Er bod y cod o dan drwydded gyfyngedig (III 1.2 Trwyddedu), cyhoeddwyd pensaernïaeth *Deep Speech* Hannun mewn papur (Hannun, et al., 2014). Mae'r prawf cysyniad gaiff ei amlinellu yn Hannun et al. (2014) yn dangos model syml fyddai'n gallu cystadlu gyda'r modelau gorau yn y byd; cyfrifwyd

WER o 16.0% (2014, tt. 1, 8). Mae'n defnyddio set brofi o 300 awr a set hyfforddi o 2,000 awr Saesneg ar gyfer hyfforddi'r model. Datblygodd hyn i *Deep Speech 2* (Amodei, et al., 2016) oedd yn ei ehangu i fod yn beiriant amlieithog, dal heb angen gwybodaeth ieithyddol arbenigol. Yr iteriad (*iteration*) nesaf oedd *Deep Speech 3* oedd yn ei ehangu eto (Battenberg, et al., 2017). Er hyn, dylid tynnu sylw at sylwadau Battenberg et al. (2017) wrth iddynt nodi bod y camau ychwanegol, hynny yw gwneud i ffwrdd â modelau iaith allanol, yn gallu achosi rhai problemau gan eu bod yn colli'r cyfle i hyfforddi ar ddata parth-benodol, fyddai'n gallu gwella'r peiriant mewn rhai cyd-destunau. I ddatrys hyn, datblygwyd *Cold Fusion* gan Sriram et al. (2017), er mwyn gweithio gyda modelau syml *Deep Speech*, ond wrth ddefnyddio modelau iaith allanol, wedi'u hyfforddi'n barod.

Yn 2017, datblygwyd gweithrediad Mozilla o hyn: *DeepSpeech* (Mozilla, 2019-2020). Mae *DeepSpeech* yn beiriant cod agored (III 1.2 Trwyddedu), sy'n cynnig cyfleoedd newydd i ieithoedd llai eu hadnoddau, sydd cyn hyn heb gael mynediad ymarferol i rwydweithiau niwral pen-i-ben (II 3.3 Modelau Pen-i-ben). Gwelir ar wefan fersiwn diweddaraf *DeepSpeech* (Davis, 2019) bod fersiwn 0.5.0, yn cyfrifo 8.22% WER ar Saesneg Americanaidd wrth hyfforddi ar gyfuniad o Fisher (tua 2,000 awr) (Cieri et al., 2004), *Librispeech* (1,000 awr) (Panayotov et al., 2015) a *Switchboard* (260 awr) (Godfrey et al., 1992). Fel cyd-destun, defnyddir 5,000 awr o ddata er mwyn hyfforddi *Deep Speech* Baidu ar gyfer Saesneg (Hannun, et al., 2014, t. 6). Gwelir maint y corpws isod yn Tabl 24, yn erbyn rhai corpora poblogaidd Saesneg eraill.

Corpws	**Oriau**
WSJ	80
Switchboard	300
TED-LIUM	452
Fisher	tua 2,000
Librispeech	tua 1,000
Baidu	5,000

Tabl 24 Corpora ar gyfer hyfforddi adnabod lleferydd Saesneg

Fel nodir ym mhennod II 4 Dulliau i Ymdopi gyda Llai o Ddata, mae timau ymchwil wedi cynnal arbrofion gyda *DeepSpeech* ar gyfer ieithoedd llai eu hadnoddau

sy'n dioddef o ddiffyg data, megis Agarwal a Zesch (2020), lle daethpwyd i'r canlyniad nad oedd yn ddull ymarferol ar gyfer Almaeneg y Swistir am y tro (71.5% WER ar <100awr). Un arall yw Hjortnæs et al. (2020), lle daethpwyd i ganlyniad tebyg ar gyfer Komi, iaith Wralig caiff ei siarad yn Rwsia (99.1% WER ar 35 awr). I atgoffa, yr isaf y WER, y gorau oll yw'r canlyniad. Er hyn, mae timau ymchwil eraill wedi edrych yn ehangach ar ddefnydd technegau pen-i-ben yn gyffredinol wrth ymdopi gyda data llai. Gan ddefnyddio dulliau trosglwyddo dysgu (mwy yn II 4 Dulliau i Ymdopi gyda Llai o Ddata), cyfrifwyd canlyniadau addawol ar gyfer Twrcaidd (80 awr, 45.8% WER, (Baetev et al., 2018)), ar gyfer ieithoedd India (1590 awr, 22.91-29.05% (Toshinwal, et al., 2018)) a chasgliadau penodol o ieithoedd (3 awr, tua 60% WER (Cui, et al., 2015)). Ceir trafodaeth ehangach ar hyn yn 4.

Mae *DeepSpeech* yn beiriant pen-i-ben sy'n golygu ei fod yn dysgu sut i adnabod llythrennau yn uniongyrchol o'r data sain. Gan nad oes angen gwybodaeth ieithyddol, mae'n cynnig opsiynau symlach o ddatblygu i ieithoedd llai eu hadnoddau sydd, fel nodir uchod, yn aml yn dioddef o ddiffyg arbenigedd. Yr anfantais i hyn yw, gan nad oes gwybodaeth ieithyddol, mae'n dibynnu'n llwyr ar y data ac felly mae angen maint sylweddol mwy o'r data hwnnw, yn ogystal â phŵer cyfrifiadurol mawr i redeg y peiriant. Ar gyfer ieithoedd llai eu hadnoddau, sy'n aml yn dioddef o ddiffyg data hefyd, mae'r pwyslais ar ddata mwy yn anfantais allweddol. Dyma grynhoad o'r manteision a'r anfanteision mae *DeepSpeech* yn cynnig i ieithoedd llai eu hadnoddau:

Manteision	**Anfanteision**
Dim angen gwybodaeth ieithyddol, **arbenigol**	Angen system gyfrifiadurol **bwerus** sy'n defnyddio nifer o GPUs
Wedi **symleiddio'n** sylweddol, heb angen gwybodaeth ieithyddol arbenigol	Angen >1,000 o oriau o **ddata** amrywiol (roedd gan Mozilla fynediad at gorpora Saesneg dros 2,000 awr ac roedd dal angen mwy o ddata 'to achieve excellent results' (Morais, 2017))
Dogfennaeth a chymorth ar gael gan **gymuned** Mozilla	

Tabl 25 Manteision ac anfanteision *DeepSpeech* Mozilla

Pwrpas y profion yn y bennod hon yw cymharu canlyniadau *Kaldi* a *DeepSpeech* gan ddefnyddio'r un data Cymraeg, er mwyn asesu a oes digon o ddata addas ar gael er mwyn gwneud yn fawr o beiriant mwy diweddar *DeepSpeech*. Caiff nodweddion perthnasol *Kaldi* a *DeepSpeech* eu crynhoi yn Tabl 26 isod.

Kaldi	***DeepSpeech***
• Angen maint sylweddol o ddata (ond yn dal i fod yn effeithiol gyda diffyg data, gwelir pennod V 3 Canlyniadau) • Angen gwybodaeth ieithyddol, arbenigol • Defnyddio dulliau diweddar croesryw HMM-DNN • Cefnogaeth cymuned dechnolegol *Kaldi*	• Angen maint sylweddol **fwy** o ddata • **Dim** angen gwybodaeth ieithyddol, arbenigol • Defnyddio dulliau **mwy** diweddar DNN pen-i-ben • Cefnogaeth cymuned dechnolegol Mozilla

Tabl 26 Cymharu *Kaldi* a *DeepSpeech*

Gwelir yn Tabl 26 bod *Kaldi* a *DeepSpeech* yn cynnig manteision tebyg i'w gilydd o ran dulliau diweddar a chefnogaeth cymuned dechnolegol, ond ymddengys bod gan *DeepSpeech* dechnegau mwy diweddar sy'n cynnig manteision ychwanegol dros *Kaldi*. Y gymhariaeth bwysicaf oll yn y cyd-destun hwn yw canlyniadau *Kaldi* a *DeepSpeech* wrth ymdopi gyda diffyg data.

1.3. Cwestiynau Ymchwil a Rhagdybiaethau Perthnasol

Y cwestiwn ymchwil dan sylw yn y bennod hon yw rhif 1 ynghylch pa ffordd o fodelu'r Gymraeg yn gyfrifiadurol sy'n creu'r system adnabod lleferydd mwyaf cadarn. Mae'n edrych i sut mae peiriant *DeepSpeech* yn ymdopi gydag adnabod lleferydd Cymraeg sy'n dioddef o ddiffyg data. Fy rhagdybiaeth yw y bydd *DeepSpeech* yn rhoi canlyniadau gwaeth na *Kaldi* wrth ymdopi gyda diffyg data ar sail gwaith gan ieithoedd llai eu hadnoddau eraill sydd wedi dod i'r un casgliad (1.1), felly'n dangos bod diffyg data yn cael mwy o effaith ar y peiriant *DeepSpeech* na chit offer *Kaldi* gan nad yw'n defnyddio geiriadur ynganu.

2. Dull

Yn 2.1, ceir cyflwyniad i'r data a ddefnyddir yn y profion hyn. Defnyddir data corpws Paldaruo 4.0, *Common Voice-cy* (13/02/19), ac data llawn dilys S4C i fesur effeithiolrwydd *DeepSpeech* wrth brosesu gwahanol feintiau o gorpora Cymraeg. Yn 2.2, esbonnir pa brofion a gynhelir, y dulliau profi a ddefnyddir, yn ogystal ag adeiladwaith *DeepSpeech*.

2.1. Data

Er mwyn cynnal cymhariaeth gynhwysfawr, cynhelir 3 prawf yn defnyddio data o feintiau gwahanol. Gwelir isod yn Tabl 27, manylion y data sy'n cael ei ddefnyddio yn y profion. Lle defnyddir y data mewn profion blaenorol yn y traethawd hir hwn, defnyddir yr un fersiynau o'r corpws hwnnw. Gwelir cynnwys y data yn Atodiad B, Atodiad C ac Atodiad H. Ceir manylion y dulliau profi isod yn 2.2.

Prawf	Data	Amser (awr)	Math y Data
1	*Common Voice-cy* (13/02/19)	42	brawddegau llawn
2	*Common Voice-cy* (13/02/19) +S4C	48.02	brawddegau llawn +geiriau unigol +brawddegau llawn amrywiol
3	*Common Voice-cy* (13/02/19) +S4C +Paldaruo 4.0	86.02	brawddegau llawn +geiriau unigol +brawddegau llawn amrywiol

Tabl 27 Manylion data profion *DeepSpeech*

Dengys Tabl 16, pennod VI 3 Canlyniadau, bod *Common Voice-cy* yn perfformio'n well fel data hyfforddi a phrofi na Paldaruo 4.0 (3.44% a 4.72% WER yn ôl eu trefn (esboniad WER yn 2.2)) felly dyna ffocws prawf 1 yn y bennod hon. Er mwyn asesu effaith cynyddu maint y data, ffocws prawf 2 a 3 yw cyfuno'r corpora un ar y tro er mwyn dysgu mwy am effaith diffyg data ar *DeepSpeech*. Ym mhrawf 2, defnyddir cyfuniad o *Common Voice-cy* a data llawn dilys S4C (Tabl 19, VII 2.1 Data), ac ym mhrawf 3 ychwanegir Paldaruo 4.0 gan mai dyma oedd y fersiwn mwyaf mewn maint ar adeg profi (Tabl 10, V 2.1 Data). Prawf 3 yw'r 'prawf mawr' felly, sy'n cyfuno'r holl ddata addas oedd ar gael ar adeg profi.

Wrth gynnal y profion hyn, bydd modd dysgu mwy am effaith gwahanol feintiau corpora wrth ddefnyddio *DeepSpeech* ac asesu a yw hwn yn opsiwn addas ar gyfer y Gymraeg, lle mae diffyg data. Bydd canlyniadau'r profion hyn yn edrych i fy rhagdybiaeth uchod (1.3): nad oes digon o ddata addas ar gael eto er mwyn gallu cymryd mantais o ddatblygiadau *DeepSpeech* yn y Gymraeg am y tro.

2.2. Gweithdrefn

Gwelir uchod pa ddata gaiff ei ddefnyddio yn y profion yn y bennod hon. Yn y rhan hon, ceir manylion y profion a redir, yna esboniad o'r dulliau a ddefnyddir, y camau a ddilynir er mwyn rhedeg y profion ac yna, adeiladwaith *DeepSpeech* ei hun. Fel nodir uchod, rhedir 3 prawf yn y bennod hon er mwyn dysgu mwy am effaith diffyg data ar *DeepSpeech*. Gwelir manylion y profion isod yn Tabl 28.

Prawf	**Set Hyfforddi (90%)**	*Amser (awr)*	**Set Brofi (10%)**	*Amser (awr)*
1	*Common Voice-cy* (13/02/19)	37.8	*Common Voice-cy* (13/02/19)	4.2
2	*Common Voice-cy* (13/02/19) +S4C	42.32	*Common Voice-cy* (13/02/19) +S4C	4.70
3	*Common Voice-cy* (13/02/19) +S4C +Paldaruo 4.0	76.52	*Common Voice-cy* (13/02/19) +S4C +Paldaruo 4.0	8.50

Tabl 28 Profion *DeepSpeech*

Yn debyg i brofion pennod V 2.2 Gweithdrefn, rhennir y data profi i setiau hyfforddi (90%) a phrofi (10%). Defnyddir yr asesiad cynhenid (*instrinsic evaluation*) (Jurafsky & Martin, 2019, t. 69) er mwyn cyfrifo'r canlyniadau o'r setiau hyn. Ceir cyflwyniadau i'r dull hwn ym mhennod V 2.2 Gweithdrefn ond, i atgoffa, mae'r asesiad cynhenid yn brawf unigol sy'n neilltuo 10% o'r data er mwyn ei ddefnyddio fel set brofi. Ceir manylion y setiau hyfforddi a phrofi a ddefnyddir uchod felly, yn Tabl 28. Ceir esboniad o'r metrig WER ym mhennod V 2.2 Gweithdrefn ond, i atgoffa, mae'n cyfrifo nifer y dileadau, mewnosodiadau ac amnewidiadau, felly'r isaf yw'r WER, y gorau oll yw'r canlyniad.

Er mwyn rhedeg yr arbrofion hyn yn *DeepSpeech*, dilynir y camau yn y rhestr isod ar gyfrifiadur arbenigol sydd â cherdyn graffig gyda GPUs: NVIDIA GeForce GTX 1080 Ti, NVIDIA GeForce RTX 2080 Ti.

1. llwytho'r data dan sylw (Paldaruo 4.0, *Common Voice-cy* (13/02/19), S4C);
2. creu ffeil .csv i nodi lleoliad y ffeiliau;
3. estyn y corpws;
4. creu set hyfforddi (90%) a set brofi (10%);
5. paratoi ffeil alphabet.txt (manylion isod);
6. addasu unrhyw fanylion pellach os oes angen e.e. graddfa ddysgu (*learning rate*), maint y llwyth (*batch size*), pwysau (*weights*) neu, gyfyngu ar gyfnodau (*epochs*) (manylion isod);
7. rhedeg y prawf.

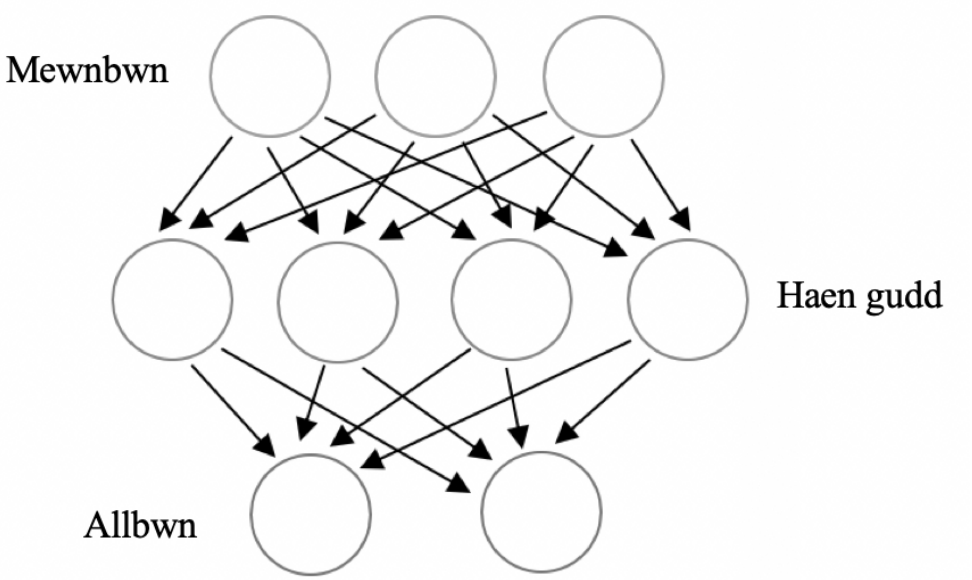
Fel yn y rhestr gyfatebol nodir ym mhennod V 2.2 Gweithdrefn, gwelir bod dal angen llwytho'r data a chreu ffeil .csv i nodi lleoliad y ffeiliau, yn ogystal ag estyn y corpws a chreu setiau hyfforddi a phrofi. Yn wahanol i'r rhestr honno, defnyddir ffeil 'alphabet.txt' yn hytrach na geiriadur ynganu. Y gwahaniaeth rhwng y ffeiliau hyn yw bod ffeil alphabet.txt yn rhestri bob llythyren gellid adnabod yn uniongyrchol o'r data, a bod geiriadur ynganu yn gasgliad o'r geiriau â'r trawsgrifiadau cyfatebol o'r seiniau hynny. Ar gyfer y ffeil alphabet.txt felly, mae angen gwirio bod pob llythyren neu symbol sy'n ymddangos yn y data wedi'i gynnwys yn y ffeil honno. Dylid nodi hefyd, bydd unrhyw resi gwag, llythrennau estron fel 'é', neu gamgymeriadau yn y ffeiliau .csv yn baglu'r ffeil alphabet.txt felly dylid gwirio hwn hefyd.

Wrth ymdopi gyda data llai, y rhagdybiaeth yw bydd y peiriant yn methu a chyrraedd yr ateb cywir o'r data yn unig, ac felly'n stopio'n gynnar. Mae'r raddfa ddysgu yn cyfeirio at y paramedr sy'n rheoli maint addasu'r model ar gyfer iteriad nesaf y prawf. Mae'r prawf yn cael ei stopio os nad oes unrhyw welliant neu waethygiad yn cyfrif fel colled (*loss*) rhwng yr iteriadau. Gellid lleihau'r raddfa ddysgu mewn rhai sefyllfaoedd, er mwyn galluogi'r prawf i barhau am hirach, a rhoi cyfle gwell i'r peiriant adnabod y lleferydd (defnyddir graddfa ddysgu o 0.0001 yn y profion hyn i ddechrau). Mae maint y llwyth yn cyfeirio at faint o ffeiliau mae'r peiriant yn ceisio prosesu ar unwaith ac felly, gellid lleihau maint y llwyth (rhagosodiad yw 1) er mwyn galluogi'r peiriant i ganolbwyntio ar lai o ffeiliau ar y tro, eto yn rhoi cyfle gwell i'r peiriant adnabod y lleferydd. Yr anfantais i hyn yw bod y prawf yn arafu. Wrth brosesu llai o ffeiliau ar yr un pryd, mae'r amser prosesu yn cynyddu ac mae'r profion yn gallu para llawer hirach cyn

cyfrifo canlyniad WER. Mae modd amrywio'r pwysau hefyd, er mwyn rhoi mwy o hyder i'r peiriant wrth ddadansoddi ac adnabod geiriau hirach a chymhleth (rhagosodiad tua 0-0.5). Mae prawf yn mynd trwy gyfnodau, sef lefelau gwahanol, cyn cyfrifo'r WER terfynol. Er mwyn mynd i'r afael â hyn, mae modd rhoi cap ar faint o gyfnodau mae'r prawf yn gallu cyrraedd cyn gorffen er mwyn osgoi rhedeg yn rhy hir. Er enghraifft, rhoddir cap o 1,000 cyfnod ar y profion yn y traethawd hir hwn (rhagosodiad), felly ni fyddai modd i'r profion barhau dros 1,000 cyfnod. Os yw'r prawf yn stopio ar ôl nifer bach o gyfnodau, mae'n awgrymu nad yw'r data yn ddigon cryf i hyfforddi'n llwyddiannus.

Adeiladwaith Peiriant DeepSpeech

Fel esbonnir ym mhennod V 2.2 Gweithdrefn, prif rôl cit offer yw hyfforddi modelau sy'n cyfrannu at y broses adnabod lleferydd, ynghyd â chydrannau eraill sy'n paratoi'r wybodaeth ieithyddol. Mae'r bennod honno yn nodi bod *Kaldi* yn git offer croesryw sy'n defnyddio modelau HMM-DNN, yn cynnig manteision dros fodelau HMM-GMM. Mae *DeepSpeech* yn beiriant DNN pen-i-ben yn llwyr, sy'n cyfuno'r holl gydrannau hyfforddi. Mae hyn yn symleiddio'r broses hyfforddi yn sylweddol (Hannun, et al., 2014). Mae'n hyfforddi drwy ddefnyddio modelau dilyniant-i-ddilyniant (*sequence-to-sequence*). Gwelir manylion pellach mewn cymhariaeth o ddulliau pen-i-ben yn gyffredinol yn Prabnhavalkar (2017) lle defnyddir tua 12,500 awr o ddata, ac ymddengys bod peiriant pen-i-ben yn gystadleuol â dulliau traddodiadol mewn rhai tasgau, ond nid pob un. Yn hytrach na defnyddio gwybodaeth ieithyddol, mae'r peiriant yn cynhyrchu dilyniant ac yn cyfrifo canlyniad amcangyfrifol ar sail y data. Yn lle pensaernïaeth cit offer (V 2.2 Gweithdrefn, Ffigwr 16), sy'n dangos modelau'n cyfrannu at y broses dadgodio, mae adeiladwaith *DeepSpeech* yn gyfres o haenau cudd rhwng y mewnbwn a'r allbwn, yn cyfuno'r cydrannau hyn i gyd. Gwelir pensaernïaeth model DNN yn fras isod yn Ffigwr 20.



Ffigwr 20 Pensaernïaeth model DNN yn fras

(wedi addasu o Meyer (2019, t. 45))

Mae hyn yn golygu bod y peiriant yn gallu cyfrifo canlyniadau gyda llai o waith paratoi gan y datblygwr. Fel nodir uchod yn 1.2, rhaid ystyried rhai rhagangenrheidiau anfanteisiol hefyd. I atgoffa, er mwyn hyfforddi DNN pen-i-ben yn llwyddiannus, mae angen:

1. system gyfrifiadurol bwerus sy'n defnyddio nifer o GPUs;
2. miloedd o oriau o ddata amrywiol (Hannun, et al., 2014).

Fel nodir uchod yn 1.2, ac ym mhennod II 4 Dulliau i Ymdopi gyda Llai o Ddata, dengys gwaith timau ymchwil megis Baetev et al. (2018) (80 awr), Toshinwal et al. (2018) (1500 awr), a Cui et al. (2015) (3 awr), bod modd cyfrifo canlyniadau addawol gyda data llai na chyfanswm 5,000 awr Hannun et al. (2014). Er y canlyniadau addawol hyn, dylid nodi eto eu bod yn defnyddio dulliau trosglwyddo dysgu; ceir trafodaeth ehangach am oblygiadau hyn yn 4.

3. Canlyniadau

Yn y rhan hon, ceir canlyniadau WER y profion yn cymharu effaith diffyg data ar git offer *Kaldi* a pheiriant *DeepSpeech*. Gwneir hyn oherwydd bod *DeepSpeech* yn cynnig nifer o fanteision datblygu dros *Kaldi* fel nodir yn 1.2. Mae *DeepSpeech* yn gwneud i ffwrdd â gwybodaeth ieithyddol (geiriadur ynganu yn yr achos hwn), sy'n gallu bod yn heriol i greu dan amgylchiadau llai eu hadnoddau gan fod angen comisiynu gwaith arbenigol. Er mwyn ymdopi â hyn, mae angen maint mwy o ddata ar *DeepSpeech* felly cynhelir y profion hyn er mwyn darganfod os yw'r maint o ddata addas sydd ar gael ar gyfer y Gymraeg yn ddigonol ar gyfer *DeepSpeech*.

Nodir yn 2.1 bod y profion hyn yn defnyddio *Common Voice-cy* (42 awr), data llawn dilys S4C (6.02 awr) a Paldaruo 4.0 (38 awr). Defnyddir canlyniadau model ystadegol tri3b *Kaldi* gan fod Tabl 12 yn dangos bod y model hwn yn fwy effeithiol na model niwral nnet2 *Kaldi* wrth ymdopi gyda'r data dan sylw. Mae esboniad o'r metrig WER ym mhennod V 2.2 Gweithdrefn ac uchod yn 2.2 ond i atgoffa, yr isaf yw'r WER, y gorau oll yw'r canlyniad. Defnyddir yr asesiad cynhenid (*instrinsic evaluation*) i gyfrifo'r canlyniadau fel nodir yn 2.2 felly nid oes cyfrifiad o ystadegau disgrifiadol fel ym mhenodau blaenorol y traethawd hir hwn.

Prawf	**Data**	**Amser (awr)**	***Kaldi* (tri3b) (WER%)**	***DeepSpeech* (WER%)**
1	*Common Voice-cy* (13/02/19)	42	3.44	99.80 (21 cyfnod)
2	*Common Voice-cy* (13/02/19) + S4C	47.02	56.55	98.00 (7 cyfnod)
3	*Common Voice-cy* (13/02/19) + S4C + Paldaruo 4.0	85.02	28.37	99.80 (74 cyfnod)

Tabl 29 Cymharu *Kaldi* gyda *DeepSpeech* (data *Common Voice-cy* (13/02/19), data llawn dilys S4C, a Paldaruo 4.0)

Mae Tabl 29 yn dangos canlyniadau adnabod lleferydd Cymraeg wrth ddefnyddio meintiau gwahanol corpora gyda *Kaldi* ac yna, gyda *DeepSpeech*. Fel nodir uchod, lle defnyddir y data mewn profion blaenorol yn y traethawd hir hwn, defnyddir yr un fersiynau yn y bennod hon. Wrth gymharu canlyniadau Tabl 29, gwelir bod *Kaldi* yn rhoi

canlyniadau WER is na *DeepSpeech*, sy'n golygu ei fod yn rhoi canlyniadau gwell. Mae'r gwahaniaeth rhwng y canlyniadau *Kaldi* a *DeepSpeech* yn fawr, sy'n awgrymu'n gryf bod *Kaldi* yn ymdopi'n well gyda diffyg data na *DeepSpeech* (ail-rhedwyd y profion hyn er mwyn dilysu'r gwahaniaeth mawr hwn). Gwelir nad oes gwahaniaeth rhwng WER *DeepSpeech* prawf 1 a 3, lle ychwanegwyd 43.02 awr, yn awgrymu nad yw ychwanegu mwy o ddata yn effeithio'r canlyniadau ar y niferoedd bach hyn. Er bod prawf 2 yn dangos WER ychydig yn is, nodir nad yw'r prawf yn cyrraedd ymhellach na 7 cyfnod, sy'n isel iawn. Fel nodir uchod yn 2.2, os yw nifer y cyfnodau'n isel iawn, mae'n awgrymu bod problem gyda'r data.

Yn ychwanegol i hyn, ymddengys bod *Kaldi* yn ymateb yn gyflymach i newidiadau ym maint y data gan fod canlyniadau *Kaldi* yn amrywio ar gyfartaledd o 40.65 lle dim ond cyfartaledd o 1.80 o newid sydd rhwng canlyniadau *DeepSpeech*. Gan fod WER *DeepSpeech* mor uchel, ac yn gyson er gwaethaf newidiadau ym maint y data, ymddengys bod maint y data mwyaf sydd ar gael yn y profion hyn (85.02 awr) ymhell o lefel digonol data *DeepSpeech*. Wrth ystyried nifer y cyfnodau hefyd, mae profion *DeepSpeech* yn cyrraedd cyfartaledd o 34 cyfnod cyn stopio, sy'n isel iawn o'i gymharu gyda'r mwyafswm o 1,000 cyfnod posib. Fel nodir yn 2.2, dyma gyfanswm y lefelau gwahanol gall y peiriant gyrraedd cyn cyfrifo'r WER terfynol. Wrth arbrofi ymhellach gyda'r set fwyaf (prawf 3) i sicrhau nad oes modd gwella'r WER wrth amrywio rhywbeth heblaw maint y data, arbrofir wrth leihau'r raddfa ddysgu ymhellach, o 0.0001 i 0.00001. Fodd bynnag, mae'r canlyniadau'n aros yr un peth fel gwelir isod yn Tabl 30. Fel nodir yn 2.2, mae'r raddfa ddysgu yn dynodi beth yw maint addasu'r model ar gyfer iteriad nesaf y prawf. Os yw'r raddfa ddysgu yn 0.0001, bydd y prawf yn stopio os nad oes gwelliant neu golled rhwng yr iteriadau ar y raddfa honno. Arbrofir ymhellach eto drwy amrywio'r pwysau yn y peiriant. Fe nodir yn 2.2, mae cynyddu'r pwysau yn gallu arwain at allu i adnabod geiriau hirach, mwy cymhleth. Nid yw'r canlyniadau hyn yn dangos newid mawr chwaith fel gwelir yn Tabl 30.

Prawf 3: *Common Voice-cy* (13/02/19) + S4C + Paldaruo 4.0 (85.62 awr)		
Graddfa ddysgu	0.0001	0.00001
Pwysau	1	2
WER%	99.80 (74 cyfnod)	99.80 (74 cyfnod)

Tabl 30 Arbrofion graddfa ddysgu a phwysau

Er bod ychydig o wahaniaeth rhwng canlyniadau'r gwahanol brofion yn Tabl 29, does dim newid arwyddocaol ac mae'r WER yn uchel iawn drwy gydol (tua 99% WER). Dylid nodi hefyd, bod canlyniadau *Kaldi* Tabl 29 yn cefnogi canlyniadau *Kaldi* Tabl 21, pennod VII 3 Canlyniadau, sy'n dangos bod ychwanegu data ar ffurf geiriau unigol (Paldaruo 4.0) neu o arddull amrywiol (S4C) yn gwaethygu'r WER. Adleisir yr angen felly i ystyried effaith modelau iaith a'r math o ddata sydd yn cael ei gasglu. Gweler pennod VII 4 Trafodaeth am fanylion pellach a goblygiadau'r ystyriaeth hon.

I grynhoi, gwelir yn Tabl 29 bod *DeepSpeech* yn cyfrifo WER uchel iawn (tua 99% WER) wrth brosesu data Cymraeg. Mae hwn yn ganlyniad llawer yn fwy na chanlyniad gorau *Kaldi* (3.44% WER). O fewn hynny, gwelir nad oes gwahaniaeth i'w weld yng nghanlyniadau *DeepSpeech* wrth ychwanegu meintiau mwy o ddata sy'n awgrymu bod maint y data ymhell i ffwrdd o lefel digonol data *DeepSpeech* am y tro.

4. Trafodaeth

Yn y rhan hon, ceir trafodaeth o ganlyniadau 3. Wrth edrych ar ganlyniadau Tabl 29 uchod, awgrymir bod *Kaldi* yn rhoi WER lawer gwell na *DeepSpeech* wrth brosesu data Cymraeg. Ymddengys fod ychwanegu meintiau gwahanol o'r data sydd ar gael, yn ogystal ag amrywio'r raddfa ddysgu a phwysau'r peiriant, yn aneffeithiol yn *DeepSpeech* hefyd sy'n awgrymu bod hwn ymhell o lefelau digonol data *DeepSpeech* wrth ddefnyddio'r dull hwn. Wrth edrych yn fanylach ar ganlyniadau *Kaldi*, gwelir eu bod yn cefnogi canlyniadau Tabl 21, pennod VII 3 Canlyniadau, sy'n dangos bod ychwanegu data ar ffurf geiriau unigol (Paldaruo 4.0) neu o arddull amrywiol (S4C), yn gwaethygu WER brawddegau *Common Voice-cy*. Gweler pennod VII 4 Trafodaeth am fanylion pellach am oblygiadau hyn wrth ystyried *Kaldi* yn y dyfodol.

Ni ddylid diystyru *DeepSpeech* yn llwyr gan ei fod yn cynnig nifer o fanteision dilys dros *Kaldi* fel nodir yn Tabl 26. I atgoffa, mae *DeepSpeech* yn cynnig:

- opsiwn lle nad oes angen gwybodaeth ieithyddol arbenigol megis geiriadur ynganu;
- system wedi'i symleiddio oherwydd hyn;
- dogfennaeth a chymorth cymuned dechnolegol Mozilla.

Wrth ddefnyddio platfform casglu data Mozilla, *Common Voice-cy* er mwyn parhau i gasglu data newydd, bydd corpws yn tyfu a bydd cyfleoedd newydd yn ymddangos i wneud yn fawr o ddatblygiadau *DeepSpeech*.

Fel nodir uchod yn 2.2, defnyddir 5,000 awr o ddata er mwyn hyfforddi *Deep Speech* Baidu ar gyfer Saesneg (Hannun, et al., 2014, t. 6). Nodir bod gwaith gan ymchwilwyr eraill (Baetev, Korenevsky, Medennikov, & Zatvornitskiy, 2018; Toshinwal, et al., 2018; Cui, et al., 2015) wedi dangos canlyniadau addawol gan ddefnyddio data llai ond gan ddefnyddio dulliau trosglwyddo dysgu. Tua'r un adeg profi a'r traethawd hir hwn, rhedir profion tebyg gyda nifer o ieithoedd gan gynnwys y Gymraeg, a gyhoeddwyd yn nhraethawd hir Meyer (2019). Yng ngwaith Meyer, gwelir bod modd arbrofi gyda throsglwyddo dysgu o fewn *DeepSpeech*, techneg lle defnyddir data o iaith (neu ieithoedd) mwy ei hadnoddau er mwyn cyflawni hyfforddi cychwynnol ar system adnabod lleferydd ac yna, ychwanegu'r iaith lai ei hadnoddau dan sylw er mwyn cwblhau'r broses hyfforddi. Yng ngwaith Meyer (2019), defnyddiodd ddata o gorpws

Common Voice-cy ar gyfer arbrofi gyda'r Gymraeg (1235 (10% set ddatblygu-manylion ym mhennod V 2.2 Gweithdrefn), 1201 (10% set brofi), 9547 (80% set hyfforddi) clip sain, a 51, 153, 75 siaradwr unigryw yn ôl eu trefn). Wrth ystyried haenau cudd adeiladwaith *DeepSpeech*, copïwyd rhwng 1 i 5 haen o Saesneg yng ngwaith Meyer (2019) er mwyn hyfforddi'r peiriant, a gan wneud gwahanol fân newidiadau, cyfrifwyd y CER (cyfradd llythyrennau gwallus) gorau Cymraeg wrth ddefnyddio 3 haen Saesneg (31.57% CER) a 4 haen Saesneg (28.75% CER). Fel noda Meyer (2019, t. 148), mae trosglwyddo dysgu yn cynnig opsiynau addawol ar gyfer ieithoedd llai eu hadnoddau, gan dynnu ar gorpora mwy mewn ieithoedd mwy eu hadnoddau. Mae gwaith Köhler (1996) yn dangos i ni fod modd defnyddio dulliau ystadegol er mwyn adnabod ieithoedd tebyg yn y cyd-destun yma, a gwella canlyniadau trosglwyddo dysgu wrth ddefnyddio ieithoedd seinegol debycach fel iaith hyfforddi a tharged. Mae'r gwelliant (2.0%) sy'n cael ei weld yng nghanlyniadau Köhler (1996, tt. 2198-2199) wrth ddefnyddio modelau amlieithog yn dangos bod mantais wrth ymchwilio ymhellach i botensial ieithoedd tebyg. Wrth ystyried y Gymraeg yn y cyd-destun hwn, gellid edrych i ieithoedd eraill Celtaidd er enghraifft (nodir eu tebygrwydd yn Ball & Müller (1992) a Fife (2002) er enghraifft). Ceir ystyriaeth bellach o'r trywydd ymchwil hwn yn 4 Arwyddocâd yn Nghyd-destun y Maes. Mae dull arall tebyg yn cael ei ystyried yng ngwaith Schultz & Waibel (2001) sy'n defnyddio tebygolrwydd ystadegol a thechnegau clystyru er mwyn datblygu modelau acwstig amlieithog a galluogi ieithoedd llai eu hadnoddau i ddefnyddio adnoddau ieithoedd mwy i wella effeithiolrwydd adnabod lleferydd. Mae'r dull yma yn gweld canlyniadau addawol, gan gyfrifo 19.6% WER ar system Portiwgaleg gan ddefnyddio dim ond 90 munud o ddata Portiwgaleg ac iaith ffonetegol debyg i hyfforddi. Mae angen mwy o waith yn y maes er mwyn deall mwy am drosglwyddo dysgu ac ieithoedd llai eu hadnoddau ond ymddengys ei fod yn cynnig manteision cyffrous, yn enwedig wrth ystyried ieithoedd ffonetegol debyg i'w gilydd.

I grynhoi, dengys Tabl 29 bod y canlyniadau yn cefnogi fy rhagdybiaeth gan fod arbrofi gyda meintiau gwahanol o ddata Cymraeg addas yn rhoi canlyniadau gwaeth wrth ddefnyddio *DeepSpeech* yn hytrach na *Kaldi*. Gan nad yw canlyniadau *DeepSpeech* yn amrywio rhwng meintiau gwahanol y data (1.80 o wahaniaeth rhwng y canlyniadau ar gyfartaledd), ymddengys nad yw *DeepSpeech* yn ymdopi'n dda gyda diffyg data, a bod hwn yn cael llai o effaith ar git offer *Kaldi* am fod ganddo gymorth gwybodaeth ieithyddol fel geiriadur ynganu. Fel nodir uchod, cyfrifir canlyniadau addawol ar gyfer y Gymraeg yn 2019 gan ddefnyddio data *Common Voice-cy* yn *DeepSpeech*, a'r dull

trosglwyddo dysgu (Meyer, 2019). I atgoffa, mae hwn yn ddull sy'n cynnig manteision newydd ar gyfer ieithoedd llai eu hadnoddau a dylid ei hystyried yn y dyfodol. Wrth ystyried y dull trosglwyddo dysgu, ac wrth ddefnyddio platfform *Common Voice-cy* i barhau i gasglu data addas, ni ddylid diystyru *DeepSpeech* ond dylid parhau i ddefnyddio cit offer *Kaldi* am y tro. Caiff oblygiadau hyn eu trafod ymhellach ym mhennod IX Trafodaeth.

IX. Trafodaeth

1. Cyflwyniad

Pwrpas y traethawd hir hwn oedd asesu sefyllfa bresennol adnabod lleferydd Cymraeg, fel iaith lai ei hadnoddau, a chynnig canllawiau ar gyfer sut i'w datblygu yn y dyfodol. Mae'r bennod hon yn ystyried canfyddiadau'r trafodaethau interim sydd yn y penodau blaenorol (2), wrth gydnabod unrhyw gyfyngiadau (5) neu waith pellach all arwain ymlaen o'r ymchwil hwn (6). Ceir y drafodaeth gyffredinol, gyda ffocws ar dri phrif amcan yr ymchwil yn 3:

1. sut orau i fodelu'r Gymraeg yn gyfrifiadurol ar gyfer adnabod lleferydd,
2. beth yw'r maint a math gorau o ddata i'w ddefnyddio, a
3. sut mae amrywio acen yn effeithio adnabod lleferydd,

gydag ystyriaeth o arwyddocâd hyn yng nghyd-destun y maes yn 4.

2. Crynhoad o'r Prif Gasgliadau

Citiau Offer

Pwrpas pennod V Citiau Offer oedd darganfod beth oedd effaith defnyddio modelau ystadegol a modelau croesryw, er mwyn darganfod pa ffordd sydd orau o fodelu'r Gymraeg, ac effaith ychwanegu mwy o'r un data, ar gyfer adnabod lleferydd. Defnyddiwyd citiau offer HTK a *Kaldi* er mwyn gweld pa un oedd yn gweithio'n well gyda maint cyfyngedig o ddata. Hefyd, defnyddiwyd meintiau gwahanol o ddata Paldaruo – casgliad cyfun o bromptiau geiriau unigol – er mwyn mesur effaith ychwanegu mwy o'r un data. Dengys y canlyniadau (crynhoad isod yn Tabl 31) bod *Kaldi* yn fwy effeithiol nac HTK wrth adnabod lleferydd Cymraeg (WER is o 11.52% WER ar ei fwyaf). Er mai cydrannau ystadegol *Kaldi* oedd yn cyfrifo'r canlyniadau gorau (yn hytrach na'r rhai niwral), roedd y canlyniadau hynny yn dal i fod yn well na'r rhai a gyfrifwyd gan fodelau ystadegol llwyr HTK (gwelliant o 6.82% WER ar ei fwyaf). Wrth ychwanegu meintiau mwy o ddata Paldaruo, cyrhaeddwyd pwynt lle nad oedd y WER yn gwella rhagor er gwaethaf ychwanegu meintiau mwy o ddata. Awgryma hyn nad yw ychwanegu mwy o'r un data yn gwneud gwahaniaeth mawr, a bod angen ystyried newid math y data.

Arweiniodd hwn ymlaen at y profion nesaf, oedd yn arbrofi i weld os oedd newid math y data i frawddegau llawn, yn hytrach na geiriau unigol, yn cael effaith gadarnhaol ar y canlyniadau.

Math o Ddata

Roedd pennod VI Math o Ddata yn adeiladu ar y canfyddiad nad oedd ychwanegu mwy o'r un math o ddata (geiriau unigol yn nghorpws Paldaruo) i *Kaldi* a HTK yn gwella'r WER yn fawr. Prif nod y bennod oedd asesu effaith math gwahanol o ddata. Gan fod data yn brin ar gyfer iaith lai ei hadnoddau, roedd hyn yn ystyriaeth bwysig. Fel rhan o ddatblygu corpws newydd oedd yn cynnwys promptiau ar ffurf brawddegau llawn, crëwyd promptiau newydd a'u recordio drwy blatfform newydd, *Common Voice-cy*. Defnyddiwyd cit offer *Kaldi* er mwyn cymharu effaith ychwanegu brawddegau llawn i'r data, ac fe welwyd ychydig o welliant o 1.28% WER yn y canlyniadau (crynhoad canlyniadau isod yn Tabl 31). Er nad yw hyn yn welliant fawr, mae'n awgrymu bod amrywio math y data yn ddull ymarferol o wella'r WER yn hytrach na glynu at yr un data.

Amrywio Iaith

Pwrpas pennod VII Amrywio Iaith oedd ymchwilio i sut mae amrywio iaith yn effeithio adnabod lleferydd. Defnyddir data wedi'i gyfrannu gan S4C ar gyfer creu setiau profi oedd yn cynrychioli acen y de ac acen y gogledd ar wahân. Profwyd y rhain ar y canlyniad gorau cyn belled (3.44% WER), sef *Kaldi* gyda data *Common Voice-cy* (oedd yn cynnwys promptiau unigol Paldaruo yn ogystal â brawddegau llawn). Gwelwyd ychydig o ddirywiad o 2.29% WER gyda set y gogledd wrth hyfforddi ar *Common Voice-cy* o gymharu gyda set y de ar yr un prawf (crynhoad canlyniadau isod yn Tabl 31). Fodd bynnag, fel nodir uchod ar gyfer y canlyniad math o ddata, dydy hyn ddim yn wahaniaeth mawr.

Er hyn, gwelwyd dirywiad mawr wrth ychwanegu'r data S4C yn ystod y broses hyfforddi yn *Kaldi*. Er enghraifft, gwelwyd dirywiad o 84.16% wrth gymharu canlyniadau *Common Voice-cy* yn unig (3.44% WER) a *Common Voice-cy* gyda set y gogledd (87.6% WER). Cynhaliwyd profion pellach felly, wrth arbrofi gyda mathau a meintiau gwahanol o ddata. Gwelwyd bod lleihau nifer y promptiau Paldaruo o'r data hyfforddi (cymharu *Common Voice*-cy gyda *Common Voice*-cy+Paldaruo 4.0) yn gwella'r canlyniadau ychydig (5.15% WER ar ei fwyaf), ond gwelwyd y gwelliant mwyaf

wrth ymdebygu'r data hyfforddi i'r data parth, hynny yw hyfforddi a phrofi ar ddata S4C. Wrth wneud hyn, cwympodd y WER i tua 60-70 ar ôl bod yn tua 80-90 gan brofi ar ddata Paldaruo 4.0, *Common Voice-cy*, a *Common Voice-cy*+Paldaruo 4.0. Er hyn, roedd y canlyniadau yn dal i ddangos dirywiad o 56.48% WER ar ei orau wrth gymharu â'r canlyniadau gorau (*Common Voice-cy* yn unig, 3.44% WER), felly cynhaliwyd dadansoddiad pellach i ffactorau eraill o fewn y data S4C allai fod yn achosi'r dirywiad. Wrth gategoreiddio gwallau yn y canlyniadau yn ôl amrywio cywair, enwau endidau, cyfnewid cod, a thafodiaith, darganfyddwyd bod amrywio cywair a chyfnewid cod yn cyfrannu'n helaeth at y gwallau hefyd. Awgryma hyn bod y sefyllfa yn fwy cymhleth na'r disgwyl felly, a bod ymchwil pellach i'w ystyried (isod).

DeepSpeech

Er mwyn asesu effaith defnyddio modelau pen-i-ben wrth fodelu'r Gymraeg, arbrofwyd gyda'r data hwn o fewn peiriant *DeepSpeech* ym mhennod VIII *DeepSpeech*. Defnyddiwyd cyfuniadau gwahanol o ddata Paldaruo, *Common Voice-cy*, a data llawn dilys S4C, ond gwelwyd dirywiad mawr yn y canlyniadau WER sy'n awgrymu nad yw *DeepSpeech* ddim yn ymdopi gyda'r data dan sylw (96.36% WER ar ei fwyaf, crynhoad canlyniadau isod yn Tabl 31). Wrth redeg arbrofion cyfatebol gyda modelau cit offer croesryw *Kaldi*, gwelwyd bod ychwanegu Paldaruo at gyfuniad *Common Voice-cy* ac S4C yn gwella'r canlyniadau yn fawr (28.18% WER), sy'n atgyfnerthu'r darganfyddiad uchod, bod ymdebygu'r data hyfforddi i'r data parth (hynny yw, ychwanegu data sydd â llai o amrywio yn y cyd-destun hwn, er enghraifft Paldaruo 4.0) yn ddull ymarferol o wella canlyniadau. Gyda chyfrifiadur pwerus, roedd y broses o redeg profion *DeepSpeech* yn symlach i redeg na phrofion *Kaldi*, yn enwedig wrth ystyried nad oedd angen mewnbwn arbenigol ieithyddol chwaith. Awgryma hyn na ddylid diystyru *DeepSpeech* yn llwyr. Wrth gasglu mwy o ddata ac wrth arbrofi gyda dulliau newydd, mae'n bosib bod potensial i'w ddefnyddio ar gyfer adeiladu modelau Cymraeg. Bydd trafodaeth bellach am hyn oll isod yn 3.

WER% *isaf=gorau*		CY1: MODEL						
		HTK	***Kaldi***					***DeepSpeech*** **(cyfnod)**
			ystadegol				***niwral***	
CY2: MAINT A MATH	**Pal 2.0**	14.70	4.35				8.87	/
	Pal 3.0	14.70	4.27				8.78	/
	Pal 4.0	**16.24**	**4.72**				**9.42**	/
	CV	/	**3.44**				/	99.80 (21)
	CV+S4C	/	56.55				/	98.00 (7)
	Pal 4.0+CV+S4C	/	28.37				/	**99.80 (74)**
			Pal 4.0	**CV**	**CV+Pal 4.0**	**S4C**		
CY3: AMRYWIO	**S4C de**	/	98.93	**85.31**	90.46	60.15	/	/
	S4C gogledd	/	99.73	**87.60**	90.86	60.94	/	/
	S4C de+gogledd	/	99.66	85.94	90.77	**59.91**	/	/

Tabl 31 Crynhoad y prif ganlyniadau

trwm *yn adlewyrchu'r canlyniadau o bwys i gwestiynau ymchwil y traethawd hir hwn*

CY: *Cwestiwn Ymchwil*

Pal: Paldaruo (geiriau unigol yn unig (2.0: 34 awr), (3.0: 36 awr), (4.0: 38 awr))

CV: *Common Voice-cy* (geiriau unigol Paldaruo 4.0 a brawddegau llawn (13/02/19) (42 awr))

S4C: de (*Gwaith Cartref* (2.40 awr)), gogledd (*Tipyn o Stad* (2.40 awr)), de+gogledd (cyfuniad (5.20 awr))

3. Trafodaeth

Fel nodir yn 1, ceir tri phrif amcan i'r traethawd hir hwn, sef

1. sut orau i fodelu'r Gymraeg yn gyfrifiadurol ar gyfer adnabod lleferydd,
2. beth yw'r maint a math gorau o ddata i'w ddefnyddio, a
3. sut mae amrywio acen yn effeithio adnabod lleferydd.

Bydd y rhan yma o'r bennod yn trafod y canlyniadau sydd wedi'u crynhoi uchod yn Tabl 31 yng nghyd-destun yr amcanion hyn.

Modelau

Er mwyn mynd i'r afael â'r amcan cyntaf, arbrofwyd gyda chitiau offer ystadegol HTK, croesryw *Kaldi* a pheiriant pen-i-ben *DeepSpeech*. Mae'r canlyniadau sy'n dangos mai *Kaldi* yw'r cit offer mwyaf effeithiol yn awgrymu mai modelau croesryw yw'r dull gorau o fodelu'r Gymraeg yn gyfrifiadurol ar hyn o bryd. Fodd bynnag, fel nodir ym mhennod V 4 Trafodaeth, cyfrifir dau set o ganlyniadau gan *Kaldi*, tri3b a nnet2, sy'n cynrychioli rhaniad croesryw *Kaldi*, ystadegol a niwral yn eu tro. Mae'r canlyniadau hyn sy'n dangos mai cydrannau tri3b ystadegol *Kaldi* sy'n rhoi'r WER isaf (gorau), yn awgrymu nad oes digon o ddata eto er mwyn gwneud yn fawr o gydrannau niwral *Kaldi* (nnet2). Fodd bynnag, mae canlyniadau ystadegol *Kaldi* dal yn well na rhai ystadegol HTK, cit offer sy'n defnyddio dulliau llai diweddar. Gall hyn awgrymu bod defnyddio cit offer mwy modern yn gwella'r canlyniadau rhywfaint hefyd. Mae'r canlyniadau hyn yn cefnogi gwaith Hindi Sri et al. (2020) ac Upadhaya et al. (2017). Yn ôl y gweithiau hyn, sy'n canolbwyntio ar ddefnyddio *Kaldi* ar gyfer Hindi, dylid modelu â *Kaldi* gan fod tri3b yn rhoi WER isel o gymharu gyda chanlyniadau eraill *Kaldi* megis nnet2. Dylid gwneud hyn wrth barhau i ystyried potensial modelau niwral – wrth gasglu mwy o ddata er mwyn gwneud yn fawr ohonynt. Caiff hyn ei atgyfnerthu gan ganlyniadau'r traethawd hir hwn ar *DeepSpeech*, sy'n awgrymu nad oes digon o ddata sain eto er mwyn gwneud yn fawr o ddulliau pen-i-ben. Mae gwaith Hernandez et al. (2019) ar y corpws TED-LIUM (cyflwyniad ym mhennod III 1.1 Ieithoedd Llai eu Hadnoddau a Diffyg Data) yn dangos bod dyblu maint data yn fras (o 207 i 452 o ddata sain wedi'i drawsgrifio), yn gwella canlyniadau o fewn *DeepSpeech*. Awgryma hyn bod tua 400 awr yn faint boddhaol o ddata sain er mwyn gweld llwyddiant o fewn *DeepSpeech.* Gall hyn fod oherwydd nad yw adeiladwaith pen-i-ben *DeepSpeech* yn defnyddio gwybodaeth ieithyddol fel geiriadur

ynganu, sy'n cefnogi'r broses o gyfrifo tebygolrwydd mewnbwn sain wrth gysylltu seiniau â ffonemau. Heb y gefnogaeth ieithyddol hon, mae'n debygol bod mwy o ddibyniaeth ar y data ei hun, all waethygu effeithiau negyddol diffyg data ar ganlyniadau WER. Fodd bynnag, mae hyn yn ychwanegu at symlrwydd y broses datblygu a phrofi, heb ofyn am fewnbwn arbenigol ieithyddol. Gyda phŵer cyfrifiadurol addas a mwy o ddata, gallai'r diffyg galw am fewnbwn arbenigol ieithyddol gynnig cyfleoedd newydd i dimau ymchwil sydd heb fynediad at adnoddau o'r fath.

Wrth ystyried sut orau i fodelu'r Gymraeg yn gyfrifiadurol felly, golyga hyn y dylid ystyried modelu'r Gymraeg gan ddefnyddio *Kaldi*, gan gadw llygad ar botensial pen-i-ben *DeepSpeech* wrth gasglu mwy o ddata. Daethpwyd i ganlyniad tebyg ar gyfer ieithoedd fel Almaeneg y Swistir er enghraifft (Agarwal & Zesch, 2020), lle methwyd i wella'r canlyniadau ymhellach na 58.9% WER gyda data cyfyngedig o fewn *DeepSpeech*. Mae'r gwaith hwn ar Almaeneg y Swistir, yn ogystal â gwaith tebyg ar yr iaith Komi, iaith Wralig a gaiff ei siarad yn Rwsia (Hjortnæs et al., 2020), yn ystyried defnyddio dulliau trosglwyddo dysgu o fewn *DeepSpeech* er mwyn gwella'u canlyniadau. Er bod y sefyllfa yn dal i fod yn 'heriol' yn ôl Agarwal & Zesch (2020), mae Hjortnæs et al. (2020) yn pwysleisio bod dal angen ystyried potensial y dull. Maent yn nodi bod y system yn cynnig allbynnau mwy synhwyrol gan ddefnyddio trosglwyddo dysgu, er nad yw hyn wedi'i adlewyrchu yn y canlyniadau (Hjortnæs et al., 2020, t. 35). Mae'n ddull sy'n dal i fod yn newydd, sy'n cynnig cyfle i hyfforddi ar ddata mwy gan iaith arall sydd â'r adnoddau addas, cyn cwblhau'r broses hyfforddi ar yr iaith darged lai ei hadnoddau. Fe'i hystyrir ymhellach fel llwybr at ymchwil pellach posib ar gyfer y Gymraeg isod yn 6.

Maint a math o ddata

Er mwyn darganfod beth yw'r maint a math gorau o ddata i'w ddefnyddio, sef ail amcan y traethawd hir hwn, cynhaliwyd arbrofion gyda meintiau gwahanol o ddata Paldaruo, yn ogystal ag amrywio math y data sain drwy gymharu promptiau geiriau unigol yn unig (Paldaruo) a chyfuniad o eiriau unigol a brawddegau llawn (*Common-voice-cy*). Mae'r canlyniadau sy'n dangos bod ychwanegu meintiau mwy o ddata yn gyffredinol yn gwella'r WER yn cefnogi ymchwil cynt sy'n awgrymu bod mwy o ddata yn arwain at ganlyniadau gwell. Mae'r canlyniadau hyn yn atgyfnerthu casgliadau Lamel et al. (2002) a Moore (2003) bod datblygiad modelau ar gyfer adnabod lleferydd yn dibynnu ar faint mawr o ddata addas. Fel nodir uchod, mae gwaith Hernandez et al. (2019) yn awgrymu

bod angen cyrraedd cannoedd o oriau o ddata er mwyn gweld llwyddiant o fewn *DeepSpeech.* Mae'n awgrymu y byddai'n rhaid casglu tua 400 awr o ddata sain Cymraeg cyn gweld canlyniadau boddhaol gyda *DeepSpeech*. Fel gwelir ym mhennod V Citiau Offer, mae systemau adnabod lleferydd ieithoedd mwy fel Saesneg yn defnyddio cannoedd a miloedd o oriau o ddata er mwyn hyfforddi eu systemau'n llwyddiannus (Ardila, et al., 2020).

Fodd bynnag, wrth ychwanegu meintiau mwy o ddata Paldaruo yn unig – set gyfyngedig o eiriau – cyrhaeddwyd pwynt lle nad oedd y canlyniadau'n gwella er gwaethaf ychwanegu mwy o ddata. Mae'r canlyniadau sy'n dangos bod ychwanegu brawddegau llawn at y data yn osgoi'r pwynt hwn, yn awgrymu bod amrywio math y data yn gallu bod yn ddull ymarferol o wella'r WER. Gall hyn fod oherwydd bod promptiau ar ffurf brawddegau llawn yn cipio mwy o ffurfiau ac amrywiadau gwahanol y geiriau all amrywio oherwydd seiniau cyfagos. Ceir cyflwyniad i eilededdau sain er enghraifft ym mhennod IV 3.3 Tafodiaith ac Acen. Yna, caiff ei ddiffinio fel achos lle mae sain neu gyfres o seiniau yn cael ei ynganu'n wahanol mewn lleferydd cysylltiedig a gofalus. Wrth ystyried canlyniadau gwell y traethawd hir hwn wrth ddefnyddio brawddegau llawn, mae'n bosib bod y lleferydd cysylltiedig a geir mewn brawddegau llawn yn adeiladu modelau acwstig sy'n well am gipio'r amrywiadau hyn. Mae'r awgrym hwn yn cefnogi gwaith Huang et al. (2020) ar Saesneg India, bod modd cipio mwy o amrywiadau iaith er enghraifft, wrth amrywio'r data. Gwelir gwelliant yng nghanlyniadau Huang et al. (2020) wrth gynnwys mwy o amrywiadau Saesneg Indiaidd yn y data hyfforddi, yn hytrach na defnyddio Saesneg Prydeinig neu Americanaidd yn unig.

Ystyriaeth arall yw'r effaith ar y modelau iaith, lle gellid ystyried bod rhan o'r gwelliant o ganlyniad i ymdebygu'r data hyfforddi i'r data profi, neu'r data parth. Ceir cyflwyniad bras i hyn ym mhennod II 4 Dulliau i Ymdopi gyda Llai o Ddata, lle nodir y cysyniad y byddai modd gweld gwelliant wrth ymdebygu'r data hyfforddi i'r data parth. Mae hyn yn atgyfnerthu gwaith Agarwal & Zesch (2020) sy'n darganfod bod allbwn y testun yn lawer byrrach na'r mewnbwn yn wreiddiol ond, wrth ymdebygu'r mewnbwn i'r parth, gwelir gwelliant mawr yn y WER. Er nad yw'r traethawd hir hwn yn ffocysu ar yr agwedd hon yn benodol, mae'r profion ar ddata S4C yn unig yn atgyfnerthu'r cysyniad. Dengys bod modd gwella WER profion ar ddata S4C wrth ddefnyddio'r un fath o ddata (S4C yn unig felly) yn y cam hyfforddi yn ogystal â'r cam profi. Ni ddylid diystyru dull Cooper et al. (2019, tt. 3-6) o adeiladu corpws ffonolegol gytbwys (adlewyrchu seiniau mwyaf cyffredin iaith) fodd bynnag. Mae canlyniadau *Kaldi* ar gyfer maint a math uchod

yn Tabl 31 yn dangos bod ychwanegu mwy o ddata Paldaruo yn gwella canlyniadau *Common Voice-cy* & S4C. Er nad yw promptiau unigol Paldaruo yn debyg i'r data parth, mae'n ffordd o wella'r canlyniadau rywfaint gyda data cyfyng. Ar gyfer defnydd o fewn adnabod lleferydd lle bo data yn brin felly, gellid ystyried creu corpws o frawddegau llawn ffonolegol gytbwys, fel llwybr at ymchwil pellach posib. Fel gyda brawddegau llawn, mae'n bosib bod y cydbwysedd ffonolegol yn gwella gallu'r modelau acwstig o adnabod ffurfiau gwahanol rhai geiriau all gael eu hamrywio dan ddylanwad seiniau cyfagos.

Amrywio Acen

Er mwyn asesu sut mae amrywio acen yn effeithio adnabod lleferydd, arbrofwyd gyda data a gyfrannwyd gan S4C. Rhannwyd y data hwn i setiau profi de a gogledd. Wrth gymharu gyda chanlyniadau *Common Voice-cy*, y gorau cyn belled, gwelwyd canlyniadau llawer gwaeth wrth ddefnyddio data S4C yn gyffredinol. Mae hyn yn cefnogi canfyddiadau ymchwil cynt gan Misnikov (2019), bod canlyniadau adnabod lleferydd yn dioddef wrth ymdopi ag acen. Gwelodd Misnikov (2019, tt. 27-28) bod amrywio acen yn dangos gwahaniaeth arwyddocaol wrth gymharu acenion Americanaidd o fewn adnabod lleferydd. Fodd bynnag, mae canlyniadau'r traethawd hir hwn sy'n dangos nad oes gwahaniaeth mawr rhwng effaith acen de a gogledd Cymru, yn awgrymu nad yw amrywio rhwng yr acenion penodol hyn yn cael effaith mawr ar y canlyniadau eu hunain.

Roedd dirywiad mawr yng nghanlyniadau data S4C yn gyffredinol o gymharu gyda chanlyniadau *Kaldi*, yn awgrymu bod ffactorau eraill o fewn y data hwnnw sy'n achosi gwallau. Wrth ddadansoddi ymhellach, gwelir bod cywair a chyfnewid cod yn cyfrannu'n helaeth at y gwallau. Er bod ychydig yn fwy o achosion o hyn yn set y gogledd, gall hyn adlewyrchu natur y data yn hytrach na'r acen ei hun. Hynny yw, bod set y de wedi'i leoli mewn ysgol, sy'n awyrgylch mwy ffurfiol, a bod rhaglen y gogledd wedi'i leoli ar ystâd dai, sy'n fwy anffurfiol. Atgyfnerthwyd hyn gan ganlyniadau enwau endidau, sy'n uwch yn set y de, lle cyfeirir at athrawon ysgol gan ddefnyddio'u teitlau Mr, Mrs ac yn y blaen. Mae cyfnewid cod yn nodwedd gyffredin mewn iaith lafar anffurfiol hefyd, fel noda Thomas & Webb-Davies (2017, t. 48), agwedd a gaiff ei hatgyfnerthu gan Ritchie & Bhatia (2012). Maent yn nodi bod siaradwyr yn gallu newid rhwng ieithoedd am nifer o resymau, gan gynnwys gyda phwy mae'r sgwrs, am beth, lle a phryd (Ritchie & Bhatia, 2012, t. 378). Wrth ystyried y canlyniadau sy'n dangos bod cyfnewid cod a chywair anffurfiol yn achosi nifer fawr o wallau wrth adnabod lleferydd,

mae'n bosib iawn bod rhain yn cyfrannu at ganlyniadau gwael data S4C. Ystyriaeth arall yw'r cysyniad a nodir uchod am ymdebygu'r data hyfforddi i'r data parth yn gwella canlyniadau. Hynny yw, gan nad oes amrywio cywair na chyfnewid cod er enghraifft, yn ymddangos yn nata *Common Voice-cy*, mae'n bosib nad yw'r data hyfforddi yn ddigon tebyg i'r data profi (S4C), ac felly, gellir achosi'r dirywiad hwn mewn WER. Un ffordd posib a gaiff ei ystyried gan ymchwil cynt i ymdopi â chyfnewid cod er enghraifft, yw creu modelau iaith paralel ar gyfer yr ieithoedd gwahanol. Dyma a wnaed gan Luján-Mares, Martínez-Hinarejos & Alabau (2007) a greodd modelau iaith paralel Sbaeneg-iaith Valencia, a gwaith Yılmaz et al. (2016) ar gyfer Frisieg-Iseldireg. Cyfrifwyd canlyniadau gwell yn y ddwy achos wrth ddefnyddio'r modelau paralel dwyieithog. Gall hyn fod yn un llwybr posib ar gyfer ymchwil pellach, a gaiff ei hystyried isod yn 6.

Wrth ystyried data S4C yn gyffredinol, dylid pwysleisio'r uchod, sef bod y recordiadau'n dod o raglenni teledu, ac felly mae'n bosib iawn bod natur sgyrsiol y data yn achosi'r dirywiad yn y canlyniadau, yn hytrach na'r acen ei hun. Mae hyn yn tynnu ar yr hyn a gyflwynir ym mhennod IV 3.1 Cywair, lle cyfeirir at waith Amalberti et al. (1993). Yn y gwaith hwnnw, cynhaliwyd arbrawf i gymharu'r ffordd mae siaradwyr yn cyfathrebu â chyfrifiaduron a phobl eraill ac fe welwyd bod grwpiau sy'n cyfathrebu â chyfrifiaduron yn defnyddio iaith fwy gofalus, a llai sgyrsiol. Wrth dynnu ar yr hyn a gyflwynir ym mhennod II 4 Dulliau i Ymdopi gyda Llai o Ddata, ac ar waith Kleynhans & Barnard (2015), mae ymdebygu data hyfforddi a phrofi yn cael effaith gadarnhaol ar y canlyniadau. Felly, wrth gyflwyno data sgyrsiol i system sydd wedi'i hyfforddi ar ddata gofalus er enghraifft, mae'n debygol na fydd y modelau yn gallu adnabod y lleferydd. Wrth gymryd *Common Voice-cy* fel enghraifft, roedd hwn yn gasgliad o recordiadau o bobl yn darllen brawddegau llawn yn uchel, yn glir ac yn bwyllog, felly mae'n bosib bod yr ynganiad yn gliriach a mwy gofalus, yn fwy addas ar gyfer adnabod lleferydd na data sgyrsiol S4C er enghraifft. Petai angen defnyddio adnabod lleferydd ar gyfer is-deitlo rhaglenni teledu eraill er enghraifft, mae'n bosib y byddai data S4C yn fwy addas gan ei fod yn debycach i'r parth. Enghraifft arall yw rhaglenni radio, sy'n aml yn cynnwys amrywio o'r fath, fel gwelir yng ngwaith Prys (2016) yn ei waith ar gyfnewid cod yn y cyfrwng. Ar gyfer pwrpasau cynorthwywyr deallus, mae canlyniadau'r traethawd hir hwn yn dangos bod data S4C yn dirywio'r canlyniadau'n fawr. Mae hyn yn awgrymu nad yw'r data yn addas ar gyfer adnabod lleferydd, sy'n aml yn dibynnu ar ddata clir a phwyllog i adlewyrchu natur y siaradwr wrth gyfathrebu â chyfrifiadur. Gwelir hyn yng

ngwaith Koulouri et al. (2016). Golyga hyn y dylid ystyried datblygu a chasglu mwy o ddata pwrpas-benodol yn y dyfodol.

Crynhoi

Prif ganfyddiadau'r traethawd hir hwn yw bod modelau ystadegol cit offer croesryw *Kaldi* yn fwy effeithiol na rhai ystadegol HTK a phen-i-ben *DeepSpeech* wrth fodelu'r Gymraeg yn gyfrifiadurol. Mae'r canlyniadau niwral cit offer croesryw *Kaldi* a chanlyniadau pen-i-ben *DeepSpeech* a geir uchod yn dangos bod angen mwy o ddata er mwyn gwneud yn fawr o dechnegau mwyaf diweddar y maes. Mae'n dangos hefyd bod potensial i wella canlyniadau a symlrwydd datblygu yn y dyfodol wrth wneud hyn. Fodd bynnag, pwysleisir nad yw pob math o ddata yn addas a bod angen ystyried ymdebygu'r data hwnnw at y pwrpas penodol. Mae'r dirywiad a geir yng nghanlyniadau S4C yn awgrymu bod angen defnyddio data tebyg i'r parth wrth hyfforddi er mwyn gweld canlyniadau boddhaol. Wrth gymharu canlyniadau S4C a chanlyniadau Paldaruo a *Common Voice-cy*, sydd wedi'u dylunio ar gyfer adnabod lleferydd, o fewn cynorthwyydd deallus er enghraifft, gwelir bod modd gwella'r canlyniadau'n fawr wrth gynllunio data pwrpas-benodol. Ceir potensial yma ar gyfer ymchwil pellach posib, wrth ystyried dylunio mwy o bromptiau ar ffurf brawddegau llawn fel *Common Voice-cy*, o bosib gan ystyried dull ffonoleg gytbwys fel a geir yng nghorpws Paldaruo, er mwyn gwneud yn fawr o ddata pan fo'n brin. Er bod canlyniadau'r traethawd hir hwn yn cefnogi ymchwil cynt bod mwy o ddata yn well, mae'n adeiladu ar y canfyddiadau hynny, nad yw unrhyw ddata yn well. Mae canlyniadau'r traethawd hir hwn yn awgrymu bod modd modelu'r Gymraeg yn effeithiol gan hyfforddi ar ddata addas, sy'n debyg i'r parth o ran ffurfiau ac amrywiadau gwahanol, ac wedi'i gynllunio ar gyfer pwrpas penodol i adlewyrchu hyn.

4. Arwyddocâd yng Nghyd-destun y Maes

Cafodd yr ymchwil yn y traethawd hir hwn ei sbarduno gan ddiffyg gwybodaeth am sut orau i barhau i ddatblygu adnabod lleferydd Cymraeg yn y dyfodol fel iaith lai ei hadnoddau. Gan mai hwn yw'r cyntaf o'i fath i ymchwilio'n benodol i fodelau, maint a math o ddata, ac effaith acen, mae gan ganfyddiadau'r traethawd hir hwn arwyddocâd o fewn adnabod lleferydd Cymraeg a'r gymuned ehangach ieithoedd llai eu hadnoddau.

Mae dyfeisiau adnabod lleferydd fel cynorthwywyr personol deallus Apple Siri, Amazon Alexa, a Google Home, yn dibynnu'n drwm ar gorpora preifat cwmnïoedd masnachol mawr. Ar gyfer datblygu adnabod lleferydd Saesneg tu allan i'r cwmnïoedd hyn, mae modd defnyddio corpora agored megis TED-LIUM (Hernandez et al., 2019) neu *Librispeech* (Panayotov et al., 2015), sy'n gyfanswm o tua 1,500 awr o ddata. Mae corpora eraill ar gael hefyd, sy'n gofyn am adnoddau cyllid er mwyn cael mynediad. Defnyddiwyd dros 2,000 awr o ddata Saesneg i hyfforddi *DeepSpeech* er enghraifft, lle'r oedd dal angen mwy o ddata '*to achieve excellent results*' (Morais, 2017), a dros 5,000 awr o ddata i hyfforddi *Deep Speech* Baidu (Hannun, et al., 2014, t. 6) yn Saesneg. Oherwydd argaeledd y data, mae ymchwil ym maes adnabod lleferydd yn aml wedi'i seilio ar Saesneg, ar gyfer Saesneg. Gwelir gwaith Jurafsky & Martin (2019, t. 502), a gwaith (Yu & Deng, 2015, t. ix) a Wang et al. (2019) er enghraifft, oll wedi seilio ar Saesneg. Mae data addas yn brin ar gyfer y Gymraeg fodd bynnag, gyda dim ond 80 awr o ddata ar gael o Paldaruo 4.0 a *Common Voice-cy* (13/02/19). Mae modd edrych i ffynonellau sy'n bodoli'n barod er mwyn casglu data ond mae costau a chyfyngiadau trwyddedu yn gallu bod yn rhwystredig. Er bod gwaith datblygu cynnar i'w weld mewn projectau technoleg lleferydd wedi'u harwain gan Williams (1994), (1995), (1999), Jones et al. (1998), (2000), Williams et al. (2006), Prys et al. (2004, t. 68) ac eraill, mae'r diffyg data addas yn golygu bod gwaith ymchwil yn y Gymraeg yn brin o'i gymharu â Saesneg er enghraifft. Gwelir gwaith datblygu mwy diweddar gyda ffocws pellach ar adnabod lleferydd yn Prys & Jones (2018) ac eraill, ond y traethawd hir hwn yw'r gwaith ymchwil cyntaf sy'n cymryd cam yn ôl i asesu gan gymharu modelau a data gwahanol er mwyn meintioli effaith gwahanol gydrannau, ac adeiladu'r system fwyaf cadarn posib. Mae canfyddiadau'r traethawd hir hwn yn cynnig canllawiau i ddatblygwyr adnabod lleferydd Cymraeg ac ieithyddion, sy'n awgrymu casglu mwy o ddata wrth sicrhau ei fod yn addas ar gyfer y pwrpas. Er enghraifft, sicrhau ei fod yn iaith gofalus, gysylltiedig er mwyn adlewyrchu adnabod lleferydd o fewn cynorthwyydd personol deallus.

Er bod ymchwil yn bodoli ar gyfer adnabod lleferydd mewn ieithoedd llai eu hadnoddau eraill, mae'r gwaith hwn yn brin, ac yn aml yn cyfrifo canlyniadau gwaeth na rhai cyfatebol yn Saesneg er enghraifft. Gwelir gwaith Goodarzi & Almasganj (2016) sy'n cyfrifo canlyniadau addawol ar gyfer adnabod lleferydd iaith Persia, wrth ddefnyddio modelau ystadegol, yn debyg i waith y traethawd hir hwn ar fodelau, ond eto yn pwysleisio bod angen mwy o ddata ac ymchwil pellach. Adlewyrchir yr angen am fwy o ddata yng ngwaith gan Bang et al. (2020) ar iaith Korea, a gan Bansal et al. (2008) ar gyfer Hindi hefyd. Gwelir gwaith ar fodelau yn Upadhyaya et al. ar gyfer Hindi eto (2017), a Punjabi (2018), lle cyfrifwyd canlyniadau addawol fodd bynnag wrth arbrofi gyda mân newidiadau i'r modelau. Dengys hyn bod angen ystyried dulliau gwahanol o ymdopi â llai o ddata pan nad oes ffordd ymarferol o gasglu maint sylweddol mwy. Mae'r gwersi a ddysgwyd o ganfyddiadau'r traethawd hir hwn yn medru cael eu trosglwyddo i ieithoedd eraill, llai eu hadnoddau, yn rhyngwladol. Gwnaed ymdrech wrth ffocysu ar wahanol gydrannau o'r system Gymraeg, i astudio ffactorau trosglwyddadwy i ieithoedd eraill, megis modelau penodol, ac amrywio iaith er enghraifft.

Mae'r systemau gorau Saesneg yn methu wrth adnabod rhai amrywiadau o Saesneg hyd yn oed, fel dangosa Misnikov (2019) yn ei astudiaeth o effaith acenion Saesneg America ar gynorthwywyr personol deallus ar y farchnad. Mae gwaith Huang et al. (2020) ar acen Indiaidd Saesneg yn adlewyrchu hyn hefyd, gan ddangos bod hwn yn her sy'n wynebu ieithoedd ar draws y byd, ddim ond yng Nghymru. Mae amrywio yn yr iaith yn her sy'n wynebu nifer o ieithoedd llai eu hadnoddau yn gyffredin (Abulimiti & Schultz, 2020). Gwelir enghreifftiau pellach ar gyfer iaith Hwngari gan Tarján et al. (2019), Twrceg gan Sak et al. (2009), yr iaith Amharic gan Pellegrini & Lamel (2009) a Tachbelie et al. (2014). Mae modd ystyried hyn yng nghyd-destun trosglwyddo dysgu sy'n ddull newydd lle defnyddir data o iaith (neu ieithoedd) fwy ei hadnoddau er mwyn hyfforddi system adnabod lleferydd cyn ychwanegu iaith lai ei hadnoddau ar ddiwedd y broses, yn manteisio ar ddata yr iaith fwy (II 4 Dulliau i Ymdopi gyda Llai o Ddata). Ceir ystyriaeth o'r dull trosglwyddo dysgu yng ngwaith diweddar Meyer (2019, t. 148). Dengys gwaith Abad et al. (2020) bod modd gweld gwelliant boddhaol wrth ddefnyddio iaith fwy ei hadnoddau yn gyffredinol, ond bod gwelliant llawer mwy wrth ddefnyddio iaith debyg er mwyn hyfforddi. Mae hyn yn codi'r cwestiwn felly, o fanteision posib i ieithoedd llai eu hadnoddau tebyg i'r Gymraeg yn arbennig, fel yr ieithoedd Celtaidd eraill. Wrth gryfhau adnoddau'r Gymraeg, mae'n cynnig opsiynau newydd i ieithoedd eraill tebyg allu tynnu arnynt, a gwella ieithoedd eraill yn eu tro. Mae canfyddiadau'r

traethawd hir hwn felly yn cynnig canllawiau trosglwyddadwy i ieithoedd eraill, llai eu hadnoddau er enghraifft, er mwyn cryfhau systemau adnabod lleferydd tu allan i Gymru, yn ogystal â'r Gymraeg hefyd:

Sut i oresgyn diffyg adnoddau drwy ddeall:

- pa fodelau sydd orau i'w defnyddio;
- pa systemau sydd orau i'w defnyddio;
- pa fath o ddata sydd orau i'w ddefnyddio neu i'w ddatblygu er mwyn sicrhau'r effeithiolrwydd mwyaf posib wrth brosesu iaith;
- sut i fanteisio ar yr hyn sydd yn barod ar gael ond hefyd,
- sut i fynd ati i ddatblygu adnoddau newydd lle bo modd.

Gwelir er enghraifft, bod modelau croesryw *Kaldi* yn ffordd o gael canlyniadau addawol pan fo data yn brin, a bod modd gwella effeithiolrwydd y peiriant drwy ystyried amrywio math y data i gynnwys mwy o enghreifftiau perthnasol a chyd-destun ffonolegol. Ceir cyflwyniad i blatfform *Common Voice* hefyd, sy'n adnodd hollbwysig wrth ystyried dyfodol technolegol ieithoedd llai eu hadnoddau eraill, ond hefyd ceir cyflwyniad i ddulliau datblygu Paldaruo, fel esiampl o gasglu data newydd mewn ffordd effeithlon a chost effiethiol.

Mae'r ymchwil a geir yn y traethawd hir hwn hefyd yn cefnogi nod Llywodraeth Cymru o greu seilwaith cadarn ar gyfer technoleg Cymraeg er mwyn gweld yr iaith a'i siaradwyr yn ffynnu (*Iaith fyw: iaith byw*, 2012; C*ymraeg 2050: Miliwn o siaradwyr Cymraeg*, 2017; *Cynllun gweithredu technoleg Cymraeg*, 2018). Wrth i'r byd technolegol newid a datblygu, maent yn cydnabod pwysigrwydd sicrhau bod technoleg Cymraeg yn ei le, i gefnogi ei siaradwyr, ac annog siaradwyr newydd hefyd. Ar gyfer strategaethau a pholisi i ddod, gall canfyddiadau'r traethawd hir hwn fod yn adnodd defnyddiol ar gyfer cyfleu gwybodaeth, awgrymiadau a chanllawiau am sut i barhau i ddatblygu, er mwyn asesu pa gymorth gall fod angen yn y dyfodol. Gwelir enghraifft bwysig o hyn ar waith yr Uned Technolegau Iaith yn y project Lleisiwr (2018), a noddwyd gan grant o Lywodraeth Cymru, lle mae cleifion yn recordio'u llais er mwyn cael ei ail-greu yn synthetig cyn iddynt eu colli o gyflyrau megis Clefyd Niwronau Echddygol neu Ganser y Gwddf. Ceir ymchwil ar gyfer Saesneg sy'n dangos bod manteision wrth ddefnyddio'r dechnoleg ar gyfer trin a monitro cleifion sy'n dioddef o ddementia (Mulvenna, et al., 2017), neu Parkinson (Jilbab et al., 2019) er enghraifft, a hynny mewn modd cost effeithiol (Zhang &

Hansen, 2019). Mae'r traethawd hir hwn yn cyfrannu felly at yr wybodaeth sydd wrth law i sefydliadau fel Llywodraeth Cymru, am sut orau i fuddsoddi yn y maes er mwyn gallu cymryd mantais o'r datblygiadau hyn.

Ceir pwyslais ar ddefnydd y dechnoleg o fewn addysg yng Nghymru mewn strategaethau yn arbennig. Gweler *Cymraeg 2050: Miliwn o siaradwyr* (2017), lle nodir bod angen 'cynyddu'r gyfran o bob grŵp blwyddyn ysgol sy'n cael eu haddysg drwy gyfrwng y Gymraeg o 22 y cant i 30 y cant', a 'gweddnewid sut rydym yn addysgu Cymraeg i bob dysgwr fel bod o leiaf 70 y cant o'r dysgwyr hynny'n gallu dweud erbyn 2050 eu bod yn gallu siarad Cymraeg' (2017, t. 12). Gellid defnyddio adnabod lleferydd i helpu cyrraedd y nod hwn, fel gwelir yng ngwaith Neri et al. (2003) a Strik et al. (2008), sy'n dangos ei fod yn ffordd ymarferol a chost effeithiol o ddysgu ail iaith. Dengys gwaith Carrier (2017) hefyd, ar ddefnydd adnabod lleferydd o fewn ysgolion Saesneg, bod modd ei ddefnyddio ar gyfer tasgau darllen yn uchel, neu gyfieithu, ac i annog hyder wrth gyfathrebu (2017, tt. 53-56). Wrth edrych ar fanteision y dechnoleg o fewn y maes addysg yn gyffredinol, daw Moussalli a Cardoso (2016) i'r canlyniad cadarnhaol ei fod yn ehangu gorwelion dosbarthiadau i ddefnyddio dulliau a deunyddiau newydd.

Dylid nodi hefyd bod manteision i ddefnyddio'r dechnoleg ym maes archifo, fel dengys gwaith Lanchantin et al. (2013), a Mohr et al. (2013), lle gwelir gwaith mynegeio recordiadau'r BBC. Dengys gwaith Jong et al. (2018) bod modd defnyddio adnabod lleferydd er mwyn echdynnu gwybodaeth (metadata) o archifau lleferydd hefyd, er mwyn categoreiddio a chwilio yn gyflymach, all fod yn fantais amserol wrth ystyried digido archifau'r BBC ac S4C. Mae'r canfyddiadau a geir yn y traethawd hir hwn yn gyfraniad unigryw felly at asesu sefyllfa adnabod lleferydd Cymraeg, a sut orau i barhau i ddatblygu, er mwyn manteisio ar ei defnydd o fewn y meysydd uchod. Mae'r Gymraeg wedi manteisio ar gefnogaeth Llywodraeth Cymru yn y gorffennol, fel gwelir yn y strategaethau uchod, ac mae'r traethawd hir hwn yn cynnig tystiolaeth ar ffurf canlyniadau arbrofion ac asesiadau o opsiynau addas, er mwyn cyfrannu at y wybodaeth angenrheidiol wrth baratoi polisi a strategaethau yn y dyfodol.

5. Cyfyngiadau

Oherwydd natur gychwynnol y gwaith hwn, mae rhai cyfyngiadau i'r ymchwil. Y cyfyngiad mwyaf amlwg yw'r diffyg data amrywiol ar gyfer astudio o fewn cyd-destun adnabod lleferydd Cymraeg. Fel gwelir uchod, dim ond 80 awr o ddata sain addas Cymraeg oedd ar gael ar y pryd a hynny wedi'i ddylunio a'i gasglu cyn deall sut byddai dylunio data gan gynnwys mwy o amrywio yn gallu gwella canlyniadau. Y traethawd hir hwn yw'r gwaith ymchwil cyntaf i feintioli'r gwelliant ar gyfer y Gymraeg. Er, nodir bod Williams (1994), (1995), yng nghanol y 1990au, wedi awgrymu dylid ystyried ehangu syntheseiddwyr i allu prosesu acen y gogledd yn y pendraw. Ffocyswyd ar acen fel man cychwyn, a arweiniodd at ystyried nifer o ffactorau eraill hefyd, fel nodir isod. Defnyddiwyd data a roddwyd gan S4C oedd angen ei rannu i acen y de a'r gogledd, gyda rhai eithriadau ar gyfer rhai cymeriadau penodol. Er mwyn sicrhau cysondeb, roedd rhaid astudio'r data yn fanwl cyn ei gyflwyno i'r system adnabod lleferydd, a thagio unrhyw acen de yn y set gogledd, a'r gwrthwyneb. Fel nodir uchod, arweiniodd y canlyniadau at ddadansoddiad pellach o rai ffactorau eraill all fod wedi cyfrannu at y canlyniadau gwael. Y ffactorau pennaf oedd cyfnewid cod a chywair anffurfiol, ac fe'u nodir isod fel llwybrau posib ar gyfer ymchwil pellach er mwyn deall mwy am effaith amrywio iaith yn y Gymraeg wrth adnabod lleferydd. Roedd yr holl waith datblygu cysylltiedig er mwyn datblygu data ar gyfer profi effaith acen yn sylweddol, a gallai hyn fod yn heriol i brojectau eraill wrth ddefnyddio data o raglenni teledu.

Cyfyngiad arall i'r gwaith hwn yw'r mesuriad o acen ei hun. Fel nodir uchod, gwnaed pob ymdrech er mwyn rhannu'r setiau data fewn i acen de a gogledd ond, wrth edrych yn fanylach ar ffactorau eraill oedd yn cyfrannu at y canlyniad gwael, daethpwyd i'r canlyniad bod y sefyllfa yn llawer mwy cymhleth. Yn y dadansoddiad pellach o ganlyniadau acen de a gogledd ar system adnabod lleferydd Cymraeg, edrychwyd am recordiadau lle'r oedd enghreifftiau o gywair anffurfiol, cyfnewid cod, enwau endidau, neu dafodiaith, yn ymddangos yn y segmentau hynny oedd wedi eu cam-adnabod. Dengys y canlyniadau bod cywair a chyfnewid cod yn cyfrannu at nifer helaeth y gwallau. Fel gwaith cychwynnol yn y maes hwn, mae'r astudiaeth ar effaith acen ar adnabod lleferydd a geir yn y traethawd hir hwn, yn arwain at lwybrau newydd ar gyfer ymchwil pellach, megis paratoi a chasglu data ffactor benodol er mwyn arbrofi o fewn system adnabod lleferydd.

6. Ymchwil Pellach

Fel nodir uchod, gan mai hwn yw'r ymchwil cyntaf o'i fath i edrych ar sut orau i fodelu'r Gymraeg yn gyfrifiadurol, pa faint a math o ddata i'w ddefnyddio, a beth yw effaith amrywiadau fel acen ar adnabod lleferydd, mae'n arwain at nifer o lwybrau ar gyfer ymchwil pellach. Yn fwyaf amlwg yw'r angen i gasglu mwy o ddata, yn gyffredinol i ddechrau, er mwyn gwella gallu modelau mwy diweddar niwral i allu ymdopi gydag adnabod lleferydd Cymraeg. Y ffordd fwyaf effeithiol ac ymarferol o wneud hyn fwy na thebyg, yw drwy ddefnyddio platfform *Common Voice-cy* gyda chymorth yr Uned Technolegau Iaith, a'r gymuned dechnoleg Cymraeg yn gyffredinol. Gwelir o'r brwdfrydedd o gwmpas *Common Voice-cy* wrth i'r platfform gael ei gyhoeddi (Prys R., 2019; Mentrau Iaith Cymru, 2019), fod y platfform yn denu sylw a chyfranwyr, a bod y system ddilysu sydd yn ei le, yn sicrhau bod y recordiadau o ansawdd da. Gwelir consensws clir ymysg datblygwyr adnabod lleferydd bod mwy o ddata yn arwain at ganlyniadau gwell, fel sydd i'w weld yn ymgyrch *Common Voice* i dyfu corpora'r byd (Ardila, et al., 2020). Ond, wrth ystyried canlyniadau'r traethawd hir hwn bod angen ystyried math y data dros y maint os yn bosib, dylid ystyried creu promptiau pellach ar ffurf brawddegau llawn er mwyn cipio natur gysylltiedig a phwyllog siaradwyr wrth gyfathrebu â chyfrifiaduron (Amalberti et al., 1993).

Fel nodir uchod yn 5, mae ymchwil y traethawd hir hwn i effaith acen ar adnabod lleferydd Cymraeg, wedi arwain at ddeall mwy am ffactorau eraill sy'n cyfrannu at y canlyniadau gwael wrth geisio prosesu acen. Yn ogystal â'r ffactorau fel acen, cywair anffurfiol, cyfnewid cod, enwau endidau, a thafodiaith, dylid ystyried a oes ffactorau eraill all fod yn cyfrannu at wallau wrth adnabod lleferydd amrywiol. Mae'n bosib bod ffactorau eraill i'w darganfod o fewn y data S4C er enghraifft tempo, goslef, traw, yn ogystal â safon y recordiadau eu hunain. Ymhellach eto, wrth gasglu data newydd, mae'n bosib bydd ffactorau ychwanegol yn codi hefyd. Wrth edrych i ffynonellau gwahanol o ddata, edrych am ffactorau eraill, a cheisio eu hynysu, mae'n bosib bydd modd deall mwy am effaith y ffactorau hyn ar adnabod lleferydd Cymraeg, a thrin modelau er mwyn gwella eu heffeithiolrwydd wrth adnabod. Gan fod canlyniadau modelau ystadegol a niwral cit offer croesryw *Kaldi* llawer yn well na modelau pen-i-ben *DeepSpeech*, mae i'w weld bod cymorth gwybodaeth ieithyddol yn allweddol er mwyn ymdopi â'r amrywiadau a'r ffurfiau gwahanol hyn. Wrth gynyddu nifer yr amrywiadau a'r ffurfiau gwahanol sy'n cael eu cynnwys yn y wybodaeth ieithyddol, mae'n bosib y byddai adeiladu modelau mwy cadarn a gwella'r canlyniadau wrth adnabod lleferydd Cymraeg.

Mae hyn yn cadarnhau felly bod angen mewnbwn arbenigwyr ieithyddol er mwyn datblygu adnabod lleferydd Cymraeg. Fel dengys gwaith Luján-Mares, Martínez-Hinarejos & Alabau (2007), ar gyflwyno cysyniad dwyieithrwydd i system adnabod lleferydd gyda model iaith ddwyieithog/modelau iaith paralel Sbaeneg-iaith Valencia, a gwaith tebyg Yılmaz et al. (2016), ar gyfnewid cod gyda model iaith ddwyieithog Frisieg-Iseldireg, gellid ystyried creu modelau ar gyfer ffactorau penodol megis model cywair ffurfiol a model cywair anffurfiol, neu fodel Cymraeg a Saesneg er mwyn ymdopi â chyfnewid cod er enghraifft.

Fel nodir uchod hefyd, mae canfyddiadau'r traethawd hir hwn ynghylch potensial modelau niwral ar gyfer y Gymraeg wedi arwain at ystyriaeth o drosglwyddo dysgu. Dyma'r dull lle defnyddir data o iaith (neu ieithoedd) mwy ei hadnoddau er mwyn cyflawni hyfforddi sylfaenol ar system adnabod lleferydd ac yna, ychwanegu'r iaith lai ei hadnoddau er mwyn gorffen hyfforddi. Mae'n golygu bod modd i'r iaith lai ei hadnoddau fanteisio ar ddata mwy yr iaith mwy ei hadnoddau er mwyn hyfforddi. Dengys gwaith cychwynnol Meyer (2019) bod modd cyfrifo canlyniadau addawol ar gyfer y Gymraeg wrth ddefnyddio trosglwyddo dysgu gyda Saesneg. Fel nodir ym mhennod VIII 4 Trafodaeth, cyfrifwyd CER (cyfradd llythrennau gwallus) o 31.57 a 28.75% CER gan Meyer (2019) wrth ddefnyddio haenau cychwynnol o ddata Saesneg cyn gorffen hyfforddi gyda'r Gymraeg o fewn peiriant *DeepSpeech*. Fel nodir ym mhennod VIII 1.2 Cymharu *Kaldi* a *DeepSpeech*, mae modelau pen-i-ben, fel a geir yn *DeepSpeech*, yn cynnig nifer o fanteision datblygu ar gyfer cymunedau llai eu hadnoddau fel y Gymraeg, felly dylid ystyried unrhyw ddulliau all arwain at eu defnydd o fewn yr ieithoedd hynny.

7. Crynodeb Cyffredinol

Mae'r ymchwil yn y traethawd hir hwn yn cynnig enghreifftiau am sut orau i fodelu'r Gymraeg yn gyfrifiadurol, beth yw effaith amrywio maint a math y data, ac effaith amrywiadau fel acen ar adnabod lleferydd, er mwyn adeiladu'r system fwyaf cadarn posib. Mae'n cynnig canllawiau i ddatblygwyr ac ieithyddion Cymraeg am sut orau i gydweithio a pharhau â'r gwaith datblygu. Mae'n bwysig oherwydd ei fod yn amlinellu gwersi trosglwyddadwy i ieithoedd llai eu hadnoddau eraill hefyd. Gall hyn arwain at ymchwil pellach i gasglu mwy o ddata amrywiol er mwyn deall mwy am amrywiadau yn y Gymraeg yng nghyd-destun adnabod lleferydd, yn ogystal ag i ddull trosglwyddo dysgu er mwyn gwneud yn fawr o ddatblygiadau niwral.

Dengys canlyniadau'r traethawd hir hwn mai defnyddio modelau *Kaldi* yw'r ffordd gorau o gael canlyniadau addawol pan fo data yn brin, er bod llwybr i ymchwil pellach wrth ystyried rôl trosglwyddo dysgu o fewn modelau niwral Cymraeg. Mae canlyniadau gwell brawddegau llawn wrth gymharu gyda geiriau unigol yn unig yn dangos bod math y data yn cael rhywfaint o effaith ar y canlyniadau. Mae'n awgrymu bod angen parhau i gynllunio data ar ffurf brawddegau llawn, o bosib oherwydd ei fod yn well am gipio'r fath o leferydd sy'n cael ei ddefnyddio wrth gyfathrebu â'r ddyfais ei hun (Amalberti et al., 1993). Mae canlyniadau llawer gwaeth data amrywiol S4C o gymharu â hyn yn dangos bod angen ystyried ymdebygu data hyfforddi ar gyfer pwrpasau penodol er mwyn gweld canlyniadau boddhaol. Mae defnydd *Kaldi* a'r angen am gynllunio data addas i gipio'r iaith dan sylw yn dangos bod y llwybr i ymchwil pellach yn un ar gyfer cydweithio rhwng datblygwyr ac ieithyddion, er mwyn deall mwy am ffactorau cyfrifiadurol ac ieithyddol sy'n effeithio'r canlyniadau.

Yn ogystal ag arwain at ymchwil pellach i ddata amrywiol ieithyddol a dulliau newydd fel trosglwyddo dysgu, mae canfyddiadau'r traethawd hir hwn yn cyfrannu at wybodaeth datblygwyr ac ieithyddion Cymraeg ac ieithoedd eraill, wrth gydnabod pwysigrwydd seilwaith technoleg ar gyfer datblygiad a goroesiad iaith.

Cyfeiriadau

Abad, A., Bell, P., Carmantini, A., & Renals, S. (2020). Cross lingual transfer learning for zero-resource domain adaptation. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (tt. 6909-6914). Barcelona: IEEE.

Abulimiti, A., & Schultz, T. (2020). Building Language Models for Morphological Rich Low-Resource Languages using Data from Related Donor Languages: the Case of Uyghur. *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), Languages Resources and Evaluation Conference (LREC 2020)* (tt. 271-276). Marseille: European Language Resources Association (ELRA).

Agarwal, A., & Zesch, T. (2020). LTS-UDE at Low-Resource Speech-to-Text Shared Task: Investigating Mozilla DeepSpeech in a low-resource setting. Yn S. Ebling, D. Tuggener, M. Hürlimann, M. Cieliebak, & M. Volk (Gol.), *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, *2624.* Zurich (held online due to COVID19 pandemic).

Amalberti, R., Carbonell, N., & Falzon, P. (1993). User Representations of Computer Systems in Human-computer Speech Interaction. *International Journal of Man-Machine Studies, 38*(4), 547-566.

Amazouz, D., Adda-Decker, M., & Lamel, L. (2017). Addressing Code-Switching in French/Algerian Arabic Speech. *INTERSPEECH 2017* (tt. 62-66). Stockholm: ISCA.

Amodei, D., Rishita Anubhai, Battenberg, E., Case, C., Casper, J., Catanzaro, B., . . . LeGres, P. (2016). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *International Conference on Machine Learning*, *48*, tt. 173-182.

Apache. (2019). *Apache*. Retrieved 06 2020, from Apache: https://www.apache.org

Apple. (2020, 02 27). *Change Siri voice or language.* Adferwyd 09 2020 o Apple: https://support.apple.com/en-gb/HT208316

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., . . . Weber, G. (2020). Common Voice: A Massively-Multilingiual Speech Corpus. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (tt. 4218-4222). Marseille: European Language Resources Association (ELRA).

Austin, P. K., & Sallabank, J. (2014). Introduction. Yn P. K. Austin, & J. Sallabank, *Endangered Languages: Beliefs and Ideologies in Language Documentation and Revitalisation* (tt. 1-26). Caergrawnt: Cambridge University Press.

Avramidis, E., & Koehn, P. (2008). Enriching Morphologically Poor Languages for Statistical Machine Translation. *Proceedings of ACL-08* (tt. 763-770). Colombus: Association for Computational Linguistics.

Awbery, G. M. (1976). *The Syntax of Welsh: A Transformational Study of the Passive.* Caergrawnt: Cambridge University Press.

Awbery, G. M. (1984). Phonotactic Constraints in Welsh. Yn M. J. Ball, & G. E. Jones, *Welsh Phonology: Selected Readings* (tt. 65-104).

Awbery, G. M. (1996). *Tafodiaith yn y Gymraeg.* Adferwyd 08 2020 o Coleg Cymraeg Cenedlaethol: https://llyfrgell.porth.ac.uk/View.aspx?id=1426~4n~NNVx9emL

Awbery, G. M. (2009). Welsh. Yn M. J. Ball, & N. Müller, *The Celtic Languages* (tt. 359-426). Llundain: Routledge.

Baetev, V., Korenevsky, M., Medennikov, I., & Zatvornitskiy, A. (2018). Exploring End-to-End Techniques for Low-Resource Speech Recognition. Yn A. Karpov, O. Jokisch, & P. Rodmonga (Gol.), *20th International Conference, SPECOM 2018*, (tt. 32-41). Leipzig.

Ball, M. J. (1984). Phonetics for Phonology. Yn M. J. Ball, & G. E. Jones, *Welsh Phonology: Selected Readings* (tt. 5-39).

Ball, M. J., & Müller, N. (1992). *Mutation in Welsh.* Llundain: Routledge.

Ball, M. J., & Williams, B. J. (2001). Speech rate variation. Yn M. J. Ball, & B. J. Williams, *Welsh Phonetics* (tt. 49-57). Lewiston: Edwin Mellen Press.

Ball, M. J., & Williams, B. J. (2001). *Welsh Phonetics.* Edwin Mellen Press.

Bang, J.-U., Kim, S.-H., & Kwon, O.-W. (2020, 03 19). Acoustic Data-Driven Subword Units Obtained through Segment Embedding and Clustering for Spontaneous Speech Recognition. *Applied Sciences, 10*(6), 2079-2093.

Bansal, P., Kant, A., Kumar, S., Sharda, A., & Gupta, S. (2008, 01). Improved hybrid model of HMM/GMM for speech recognition. *Information Technologies and Knowledge*, 69-74.

Barbosa, P. A., Camargo, Z. A., & Madureira, S. (2019). Acoustic-based tools and scripts for the automatic analysis of speech in clinical and non-clinical settings. Yn H. A. Patil, A. Neustein, & M. Kulshreshtha, *Signal and Acoustic Modeling for Speech and Communication Disorders* (tt. 69-86). De Gruyter.

Barroso, N., Ipiña, K. L., Hernández, C., Ezeiza, A., & Graña, M. (2012). Semantic speech recognition in the Basque context. *International Journal of Speech Technology, 15*, 33-47.

Battenberg, E., Chen, J., Child, R., Coates, A., Gaur, Y., Li, Y., . . . Zhu, Z. (2017, 07 24). *Exploring Neural Transducers for End-to-End Speech Recognition.* Adferwyd 07 2020 o arxiv: https://arxiv.org/abs/1707.07413

Benigo, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning, 2*(1), 1-127.

Berger, K. C., Hernaiz, A. G., Baroni, P., Hicks, D., Kruse, E., Quochi, V., . . . Soria, C. (2018). *THe DPDL Digital Language Survival Kit.* Adferwyd 10 03, 2020 o DLDP: http://www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication, 56*, 85-100.

Black, A. W., & Lenzo, K. A. (2000, 02 24). *Building Voices in the Festival Speech Synthesis System.* Adferwyd 07 2020 o Festvox: http://www.festvox.org/festvox-1.1/

Boersma, P., & Weenink, D. (2007). *Praat: doing phonetics by computer*. Adferwyd 06 2020 o Praat: doing phonetics by computer: http://www.fon.hum.uva.nl/praat/

Bohnet, B., Nivre, J., Boguslavsky, I., Farkas, R., Ginter, F., & Hajič, J. (2013, 10). Joint Morphological and Syntactic Analysis for Richly Inflected Languages. *Transactions of the Association for Computational Linguistics, 1*, 415-428.

Borsley, R. D., Tallerman, M., & Willis, D. (2007). *The Syntax of Welsh.* Caergrawnt: Cambridge University Press.

Buys, J., & Botha, J. A. (2016). Cross-Lingual Morphological Tagging for Low-Resource Languages. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (tt. 1954-1964). Berlin: Association for Computational Linguistics.

Caines, A., Bentz, C., Graham, C., Polezhl, T., & Buttery, P. (2016). Crowdsourcing a multilingual speech corpus: recording, transcription and annotation of the CrowdED corpus. Yn N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, . . . S. Piperidis (Gol.), *LREC 2016* (tt. 1-8). Portorož: ELRA.

Caon, D. R., Amerhraye, A., Razik, J., Chollet, G., Andreão, R. V., & Mokbel, C. (2011). Experiments on Acoustic Model supervised adaptation and evaluation by K-Fold Cross Validation Technique. *5th International Symposium on I/V Communications and Mobile Network, 2010*, *148*, tt. 1-5.

Carrier, M. (2017). Automated Speech Recognition in language learning: Potential models, benefits and impact. *Training, Language and Culture, 1*(1), 46-61.

Cawley, C. G., & Talbot, N. L. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research, 11*, 2079-2107.

Chan, D., & Cooper, S. (2014). *welsh-lts.* Adferwyd 07 2020 o Techiaith: https://github.com/techiaith/welsh-lts/blob/master/LICENSE.txt

Chen, X., Awadallah, A. H., & Hassan, H. (2019). Multi-Source Cross-Lingual Model Transfer: Learning What to Share. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (tt. 3098-3112). Florence: Association for Computational Linguistics.

Choukri, K., Piperidis, S., Tsiavos, P., Patrikakos, T., Gavrilidou, M., & Weitzmann, J. H. (2012, 02 24). *META-SHARE: Licenses, Legal, IPR and Licensing issues, Meta-Net Deliverable 6.3.1, February 24th, 2012.* Adferwyd 10 08, 2020 o ELRA: http://www.elra.info/media/filer_public/2015/03/30/meta-net-d613.pdf

Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K. (2004, 12 15). *Fisher English Training Speech Part 1 Speech.* Adferwyd 06 2020 o Linguistic Data Consortium: https://catalog.ldc.upenn.edu/LDC2004S13

Cieri, C., Miller, D., & Walker, K. (2004). The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. *LREC 2004* (tt. 1-3). Lisbon: ELRA.

Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., . . . Cowan, B. (2019, 09 11). The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers, 31*(4), 349-371.

Comisiynydd y Gymraeg. (2011). *Mesur y Gymraeg (Cymru) 2011.* Adferwyd 03 14, 2021 o https://www.legislation.gov.uk/mwa/2011/1/contents/enacted

Cooper, S. (2019). Speech Recognition. Yn M. J. Ball, & J. S. Damico, *The SAGE Encyclopedia of Human Communication Sciences and Disorders* (tt. 1787-1790).

Cooper, S., Chan, D., & Jones, D. B. (2018). *The Paldaruo Speech Corpus, version 1-5.* Adferwyd 07 2020 o Techiaith: http://techiaith.cymru/data/corpora/paldaruo/

Cooper, S., Jones, D. B., & Prys, D. (2014). Developing further speech recognition resources for Welsh. *Proceedings of the First Celtic Language Technology Workshop at the 25th International Conference on Computational Linguistics (COLING 2014)* (tt. 55-59). Dublin: Dublin City University and Association for Computational Linguistics.

Cooper, S., Jones, D. B., & Prys, D. (2019, 07 25). Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology. *Information, 10*(247), 1-12.

Copeland, B. J., & Proudfoot, D. (1996). On Alan Turing's Anticipation of Connectionism. *Synthese, 108*, 361-377.

Copeland, J., & Proudfoot, D. (2009). Turing's test: A Philosophical and Historical Guide. Yn R. Epstein, G. Roberts, & G. Beber, *Parsing the Turing Test* (tt. 119-138). Springer.

Costa, C. A., Pasluosta, C., Eskofier, B. M., Righi, R. D., & Bandeira, D. (2018). Internet of Health Things: Toward intelligent vital signs monitoring in hospital wards. *Artificial Intelligence in Medicine, 89*, 61-69.

Coupland, N., & Ball, M. J. (1989). Welsh and English in Contemporary Wales: Sociolinguistic Issues. *Contemporary Wales*, 7-40.

Crystal, D. (2000). *Language Death.* Caergrawnt: Cambridge University Press.

Cui, J., Kingsbury, B., Ramabhadran, B., Sethy, A., Audhkhasi, K., Cui, X., . . . Knill, K. (2015). Multilingual representations for low resource speech recognition and keyword search. *The Automatic Speech Recognition and Understanding Workshop (ASRU) 2015* (tt. 259-266). Scottsdale: IEEE.

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012, 01). Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on audio, speech, and language processing, 20*(1), 30-43.

Davies, P. (2016). Age Variation and Language Change in Welsh: Auxiliary Deletion and Possessive Constructions. Yn M. Durham, & J. Morris, *Sociolinguistics in Wales* (tt. 31-60).

Davies, P., & Deuchar, M. (2010). Using the Matrix Language Frame model to measure the extent of word-order convergence in Welsh-English bilingual speech. Yn A. Breitbarth, C. Lucas, & S. Watts, *Continuity and Change in Grammar* (tt. 77-96). John Benjamins Publishing.

Davies, P., & Deuchar, M. (2014, 03 30). Auxiliary deletion in the informal speech of Welsh-English bilinguals: A change in progress. *Lingua, 143*, 224-241.

Davis, K. (2019, 06 11). *Deep Speech 0.5.0.* Adferwyd 07 2020 o GitHub: https://github.com/mozilla/DeepSpeech/releases/tag/v0.5.0

Dechter, R. (1986). Learning while searching in constraint-satisfaction-problems. *AAAI-86 Proceedings* (tt. 178-183). AAAI.

Delič, V., Sečujski, M., Sedlar, N., Mišković, D., Mak, R., & Bojanić, M. (2014). How Speech Technologies Can Help People with Disabilities. Yn A. Ronzhin, R. Potapova, & D. Vlado (Gol.), *International Conference on Speech and Computer SPECOM 2014* (tt. 243-250). Novi Sad: Springer International Publishing.

Deuchar, M. (2014). *Documentation File for the Bangor Siarad Corpus.* Adferwyd 06 2020 o Bangor Siarad Corpus: http://bangortalk.org.uk/docs/Siarad_doc.pdf

Deuchar, M., & Davies, P. (2009). Code switching and the future of the Welsh language. *International Journal of the Sociology of Language*, 15-38.

Deuchar, M., Donnelly, K., & Piercy, C. (2016). 'Mae pobl monolingual yn minority': Factors Favouring the Pronunciation of Code Switching by Welsh-English Bilingual Speakers. Yn M. Durham, & J. Morris, *Sociolinguistics in Wales* (tt. 209-240).

Diehl, J. J. (2019). Clinical applications of speech technology for Autism Spectrum Disorder. Yn H. A. Patil, A. Neustein, & M. Kulshreshtha, *Signal and Acoustic Modeling for Speech and Communication Disorders* (tt. 161-180). De Gruyter.

Dowling, M., Lynn, T., & Way, A. (2017). A Crowd-sourcing Approach for Translations of Minority Language User-Generated Content (UGC). *The Prague Bulletin of Mathematical Linguistics*, 1-12.

Dryer, M. S. (2013). Order of Subject and Verb. Yn M. S. Dryer, & M. Haspelmath, *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dudley, H. (1939). The Vocoder. *Bell Labs Record, 17*, 122-126.

Dudley, H., Riesz, R. R., & Watkins, S. A. (1939). A Synthetic Speaker. *J. Franklin Institute, 227*, 739-764.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2020). *Ethnologue: Languages of the World. Twenty-third edition.* Adferwyd 08 2020 o Ethnologue: https://www.ethnologue.com/endangered-languages

Ellis, N. C., Roberts, D., & O'Dochartaigh, C. (2000). IT for the Welsh language: the CySill project. Yn P. W. Thomas, & J. Mathias, *Developing minority languages:*

Proceedings of the fifth international conference on minority languages (tt. 698-721). Llandysul: Gwasg Gomer.

Ellis, R. (2008). *The Study of Second Language Acquisition.* Rhydychen: Oxford University Press.

Evas, J. (2013). *Y Gymraeg yn yr oes Ddigidol.* Adferwyd 08 2020 o Cyfres Papurau Gwyn META-NET: http://www.meta-net.eu/whitepapers/e-book/welsh.pdf

Fife, J. (2002). Introduction. Yn M. J. Ball, & J. Fife, *The Celtic Languages* (tt. 3-25). Llundain: Routledge.

Gaida, C., Lange, P., Petrick, R., Proba, P., Malatawy, A., & Suendermann-Oeft, D. (2014). Comparing Open-Source Speech Recognition Toolkits. *Tech.Rep.,DHBW Stuttgart.*

Gales, M., & Young, S. (2007). The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing, 1*(3), 195-304.

Garrett, J. T., Heller, K. W., Fowler, L. P., Alberto, P. A., Fredrick, L. D., & O'Rourke, C. M. (2011, 03 01). Using Speech Recognition Software to Increase Writing Fluency for Individuals with Physical Disabilities. *Journal of Special Education Technology, 26*(1), 25-41.

Gautam, S., & Singh, L. (2019). Speech disorders in children and adults with mild-to-moderate intellectual disability. Yn H. A. Patil, A. Neustein, & M. Kulshreshtha, *Signal and Acoustic Modeling for Speech and Communication Disorders* (tt. 123-138). De Gruyter.

Ghazzali, S., Jones, D. B., & Prys, D. (2015, 10). *Tuag at Gynorthwyydd Personol Deallus Cymraeg.* Adferwyd 07 2020 o Uned Technolegau Iaith: http://techiaith.bangor.ac.uk/tuag-at-gynorthwyydd-personol-deallus-cymraeg/

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. *ICASSP 1992* (tt. 517-520). San Francisco: IEEE.

Goodarzi, M. M., & Almasganj, F. (2016). A GMM/JHMM model for reconstruction of missing speech spectral components for continuous speech recognition. *International Journal of Speech Technology, 19*, 769-777.

Gouvêa, E., & Davel, M. H. (2011). Kullback-Liebler divergence-based ASR training data selection. *Proceedings of INTERSPEECH* (tt. 2297-2300). Florence: IEEE.

Guglani, J., & Mishra, A. N. (2018, 06). Continuous Punjabi speech recognition model based on Kaldi ASR toolkit. *International Journal of Speech Technology, 21*(2), tt. 211-216.

Hallahan, W. I. (1995). DECtalk Software: Text-to-Speech Technology and Implementation. *Digital Technical Journal, 7*(4), 5-19.

Han, Z. (2004). *Fossilization in Adult Second Language Acquisition.* Clevedon: Multilingual Matters.

Hannahs, S. J. (2013). Phonological Processes. Yn S. J. Hannahs, *The Phonology of Welsh* (tt. 52-84). Rhydychen: Oxford University Press.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., . . . Ng, A. Y. (2014, 12 17). *DeepSpeech: Scaling up end-to-end speech recognition.* Adferwyd 07 2020 o Cornell University arXiv: https://arxiv.org/abs/1412.5567

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., & Estève, T. (2019, 06 13). *TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation.* Adferwyd 06 2020 o Cornell University arXiv: https://arxiv.org/pdf/1805.04699.pdf

Hicks, W. J. (2004). Welsh Proofing Tools: Making a Little NLP go a Long Way. *Proceedings of the 1st Workshop on International Proofing Tools and Language Technologies*, (tt. 1-7). Patras.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., . . . Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, 82-96.

Hjortnæs, N., Partanen, N., Rießler, M., & Francis, M. T. (2020). Towards a Speech Recognizer for Komi, and Endangered and Low-Resource Uralic Language. *Proceedings of the 6th International Workshop on Computational Linguistics of Uralic Languages* (tt. 31-37). Wien: Association for Computational Linguistics.

Hoesen, D., Satriawan, C. H., Lestari, D. P., & Khodra, M. L. (2016). Towards Robust Indonesian Speech Recognition with Spontaneous-Speech Adapted Acoustic Models. *Procedia Computer Science, 81*, 167-173.

Holmes, J. (2001). *An Introduction to Sociolinguistics.* Essex: Pearson.

HTK. (2006). *HTK License*. Retrieved 06 2020, from HTK: http://htk.eng.cam.ac.uk/docs/license.shtml

Huang, C., Chen, T., & Chang, E. (2004). Accent Issues in Large Vocabulary Continuous Speech Recognition. *International Journal of Speech Technology, 7*, 141-153.

Huang, X., Jin, X., Li, Q., & Zhang, K. (2020). On Construction of the ASR-oriented Indian English Pronunciation Dictionary. *Proceedings of the 12th Conference on Language Resources and Evaluation (ELRA 2020)* (tt. 6593-6598). Marseille: European Language Resources Association (ELRA).

Huck, M., Tamchyna, A., Bojar, O., & Fraser, A. (2017). Procuding Unseen Morphological Variants in Statistical Machine Translation. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2*, tt. 369-375. Valencia: Association for Computational Linguistics.

Humphries, J. J., & Woodland, P. C. (1997). Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition. *EUROSPEECH '97: 5th European Conference on Speech Communication and Technology* (tt. 2367-2370). Rhodes: ISCA.

ISO. (2018, 04). *ISO/TR 20694:2018 A typology of language registers.* Adferwyd 08 2020 o ISO: https://www.iso.org/standard/68852.html

Jamro, W. A. (2017). Sindhi Language Processing: A survey. *International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)* (tt. 1-8). Karachi: IEEE.

Jilbab, A., Benba, A., & Hammouch, A. (2019). Quantification system of Parkinson's disease. Yn H. A. Patil, A. Neustein, & M. Kulshreshtha, *Signal and Acoustic Modeling for Speech and Communication Disorders* (tt. 227-272). De Gruyter.

Jones, D. B., & Cooper, S. (2016). Building Intelligent Digital Assistants for Speakers of a Lesser-Resourced Language. *Proceedings of LREC 2016 Workshop: CCURL 2016 Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity* (tt. 74-79). Portorož: ELRA.

Jones, D. B., & Ghazzali, S. (2016). *Seilwaith Cyfathrebu Cymraeg.* Adferwyd 07 2020 o Uned Technolegau Iaith: http://techiaith.bangor.ac.uk/seilwaith-cyfathrebu-cymraeg/

Jones, D. B., Prys, D., Prys, M., & Prys, G. (2019, 03 13). *Llawlyfr Technolegau Iaith.* Adferwyd 07 23, 2020 o Coleg Cymraeg Cenedlaethol: https://llyfrgell.porth.ac.uk/View.aspx?id=4217~4o~orjUKoy9

Jones, D. B., Robertson, P., & Prys, G. (2015). *Gwasanaeth API Lemateiddiwr Cymraeg.* Adferwyd 07 2020 o Porth Technolegau Iaith Cenedlaethol Cymru: http://techiaith.cymru/cwmwl/api/lemmateiddiwr

Jones, D. G. (1988). Literary Welsh. Yn M. J. Ball, *The Use of Welsh* (tt. 125-171). Clevedon: Multilingual Matters.

Jones, G. E. (1984). The Distinctive Vowels and Consonants of Welsh. Yn M. J. Ball, & G. E. Jones, *Welsh phonology: selected readings* (tt. 40-64). Caerdydd: University of Wales Press.

Jones, R. J., Downey, S., & Mason, J. S. (1997). Continuous speech recognition using syllables. *Eurospeech* (tt. 1-4). Rhodes: ISCA.

Jones, R. J., Mason, J. S., Helliker, L., & Pawlewski, M. (2000). Recruitment techniques for minority language speech databases: some observations. *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, (tt. 1-5). Athens.

Jones, R. J., Mason, J. S., Jones, R. O., Helliker, L., & Pawlewski, M. (1998). SpeechDat Cymru: A Large-scale Welsh Telephony Database. *Proceedings of LREC Workshop: Language Resources for European Minority Languages* (tt. 1-7). Portorož: ELRA.

Jong, F. d., Heeren, W., Hessen, A. v., Ordelman, R., & Nijholt, A. (2018). Automated Metadata Extraction for Semantic Access to Spoken Word Archives. *International Conference on Digital Economy*, (tt. 1-10). Brest.

Juang, B. H., & Rabiner, L. R. (2004). *Automatic Speech Recognition - A Brief History of the Technology Department.* Santa Barbara: Rutgers University and the University of California.

Judge, J., Lynn, T., Ward, M., & Raghallaigh, B. Ó. (2014). Proceedings of the Celtic Language Technology Workshop (CLTW). *25th International Conference on Computational Linguistics (COLING 2014)*.

Jurafsky, D., & Martin, J. H. (2019, 10 16). *Speech and Language Processing (3rd ed. draft)*. Adferwyd 07 2020 o Stanford University: https://web.stanford.edu/~jurafsky/slp3/edbook_oct162019.pdf

Köhler, J. (1996). Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. *International Conference on Spoken Language Processing (ICSLP)* (tt. 2195-2198). Philadelphia: ISCA.

Kaldi. (2011). *Lattices in Kaldi*. Adferwyd 06 2020 o Kaldi: http://kaldi-asr.org/doc/lattices.html

Kamper, H., Mukanya, F. J., & Niesler, T. (2012, 02 09). Multi-accent acoustic modelling of South African English. *Speech Communication, 54*, 801-813.

Khelifa, M. O., Elhadj, Y. M., Abdellah, Y., & Belkasmi, M. (2017, 09 20). Constructing accurate and robust HMM/GMM models for an Arabic speech recognition system. *International Journal of Speech Technology, 20*, 937-949.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America, 82*(3), 737-793.

Kleynhans, N. T., & Barnard, E. (2015). Efficient data selection for ASR. *Language Resources & Evaluation, 49*, 327-353.

Koenecke, A., Nam, A., Lake, E., Nuddell, J., Quartey, M., Mengesha, Z., . . . Goel, S. (2020, 04 07). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(4), 1-6.

Koulouri, T., Lauria, S., & Macredie, R. D. (2016). Do (and Say) as I Say: Linguistic Adaptation in Human-Computer Dialogs. *Human-Computer Interaction, 31*(1), 59-95.

Krauwer, S. (2003). The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. *Proceedings of the 2003 International Workshop Speech and Computer* (tt. 8-15). Moscow: Springer.

Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., & Stober, S. (2017). Transfer Learning for Speech Recognition on a Budget. Yn P. Blunsom, A. Bordes, K. Cho, S. Cohen, C. Dyer, E. Grefenstette, . . . S. Yih (Gol.), *2nd ACL Workshop on Representation Learning for NLP* (tt. 168-177). Vancouver: Association for Computational Linguistics.

Labov, W. (1963). The Social Motivation of a Sound Change. *WORD, 19*(3), 273-309.

Labov, W. (2006). *The Social Stratification of English in New York City.* Washington: Centre for Applied Linguistics.

Lanchantin, P., Bell, P. J., Gales, M. J., Hain, T., Liu, X., Long, Y., . . . Woodland, P. C. (2013). Automatic Transcription of Multi-genre Media Archives. *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia* (tt. 26-31). Marseille: ISCA.

Levenshtein, V. I. (1966, 02). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics-Doklady, 10*(8), 707-710. Adferwyd o https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf

Li, L., Zhao, Y., Jiang, D., & Zhang, Y. (2013). Hybrid Deep Neural Network - Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition. *2013 Humaine*

Association Conference on Affective Computing and Intelligent Interaction (tt. 213-318). IEEE.

Llywodraeth Cymru. (2012). *Iaith fyw: iaith byw.* Adferwyd 07 2020 o Llywodraeth Cymru: https://llyw.cymru/sites/default/files/publications/2018-12/strategaeth-y-gymraeg-2012-i-2017-iaith-fyw-iaith-byw.pdf

Llywodraeth Cymru. (2017). *Cymraeg 2050: Miliwn o siaradwyr Cymraeg.* Adferwyd 08 2020 o Senedd Cymru: https://senedd.wales/laid%20documents/gen-ld11108/gen-ld11108-w.pdf

Llywodraeth Cymru. (2018, 10). *Cynllun gweithredu technoleg Cymraeg.* Adferwyd 08 2020 o Llywodraeth Cymru: https://llyw.cymru/sites/default/files/publications/2018-12/cynllun-gweithredu-ar-gyfer-technoleg-cymraeg-a-chyfryngau-digidol.pdf

Llywodraeth Cymru. (2020, 06 25). *Data am y Gymraeg a'r Arolwg Blynyddol o'r Boblogaeth: Ebrill 2019 i Fawrth 2020.* Adferwyd 10 03, 2020 o Llywodraeth Cymru: https://llyw.cymru/data-am-y-gymraeg-or-arolwg-blynyddol-or-boblogaeth-ebrill-2019-i-fawrth-2020

Luján Mares, M., Martínez-Hinarejos, C.-D., & Alabau, V. (2007). A Study on Bilingual Speech Recognition Involving a Minority Language. *Language and Technology Conference, Human Language Technology, Challenges of the Information Society*, (tt. 36-49).

Lyu, D.-C., & Lyu, R.-Y. (2008). Language Identification on Code-Switching Utterances Using Multiple Cues. *INTERSPEECH 2008* (tt. 711-714). ISCA.

Lyu, D.-C., Tan, T.-P., Chng, E.-S., & Li, H. (2015). Mandarin-English code-switching speech corpus in South-East Asia: SEAME. *Language Resources and Evaluation, 49*, 581-600.

Marchi, E., Zhang, Y., Eyben, F., Ringeval, F., & Schuller, B. (2019). Autism and speech, language, and emotion - a survey. Yn H. A. Patil, A. Neustein, & M. Kulshreshtha, *Signal and Acoustic Modeling for Speech and Communication Disorders* (tt. 139-160). De Gruyter.

Maritinet, A. (1952). Celtic lenition and Western Romance consonants. *Language, 27*, 192-217.

Matarneh, R., Maksymova, S., Lyashenko, V. V., & Belova, N. V. (2017, 09-10). Speech Recognition Systems: A Comparative Review. *IOSR Journal of Computer Engineering, 19*(5), 71-79.

Mayr, R., & Davies, H. (2011). A cross-dialectal acoustic study of the monophthongs and diphthongs of Welsh. *Journal of the International Phonetic Association, 41*(1), 1-25.

McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence.* Natick, Massachusetts, USA: A K Peters.

McCulloch, W. S., & Pitts, W. (1943, 12). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*, 115-133.

Mentrau Iaith Cymru. (2019, 06 25). *Common Voice: 22 Awr i 22 Menter.* Adferwyd 07 2020 o Mentrau Iaith Cymru: http://mentrauiaith.cymru/newyddion/common-voice-22-awr-i-22-menter/

META. (2013). *META-NET White Paper Series.* Adferwyd 10 2020 o META: http://www.meta-net.eu/whitepapers/overview

Meyer, J. (2019, 08 17). *Kaldi Troubleshooting Head-to-Toe*. Adferwyd 06 2020 o Josh Meyer's Website: http://jrmeyer.github.io/asr/2019/08/17/Kaldi-troubleshooting.html

Meyer, J. (2019). *Multi-task and transfer learning in low-resource speech recognition.* Adferwyd 07 2020 o Josh Meyer's Website: http://jrmeyer.github.io/misc/MEYER_dissertation_2019.pdf

Miesel, J., Clahsen, H., & Pienemann, M. (1981). On determining developmental stages in natural second language acquisition. *Studies in Second Language Acquisition, 3*(2), 109-135.

Mills, M. (2012, 06 01). Media and Prosthesis: The Vocoder, the Artificial Larynx, and the History of Signal Processing. *Qui Parle, 21*(1), 107-149.

Misnikov, E. (2019). *Sociophonetic factors in automatic speech recognition: a study on American English.* Adferwyd 08 2020 o Deutsche National Bibliothek: https://d-nb.info/1184789525/34

Mittal, S., & Vaishnay, S. (2019, 10). A Survey of Techniques for Optimizing Deep Learning on GPUs. *Journal of Systems Architecture, 99*, 1-45.

Modipa, T. I., Davel, M. H., & Wet, F. d. (2013). Implications of Sepedi/English code switching for ASR systems. *Proceedings of the 24th Annual Symposium of the Pattern Recognition Association of South Africa*, (tt. 64-69). Johannesburg.

Mohr, C., Saam, C., Kilgour, K., Gehring, J., Stüker, S., & Waibel, A. (2013). Slightly Supervised Adaptation of Acoustic Models on Captioned BBC Weather Forecasts.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (tt. 32-36). Marseille: ISCA.

Moore, R. K. (2001). There's no data like more data: but when will enough be enough? *Proceedings of the Institute of Acoustics*, *23*, tt. 19-26.

Moore, R. K. (2003). A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners. *EUROSPEECH*, (tt. 2582-2584). Geneva.

Morais, R. (2017, 11 29). *A Journey to <10% Word Error Rate*. Adferwyd 07 2020 o Mozilla Hacks: https://hacks.mozilla.org/2017/11/a-journey-to-10-word-error-rate/

Moussalli, S., & Cardoso, W. (2016). Are commercial 'personal robots' ready for language learning? Focus on second language speech. Yn S. Papadima-Sophocleous, L. Bradley, & S. Thouësny (Gol.), *CALL communities and culture - short papers from EUROCALL 2016*, (tt. 325-329).

Mozilla. (2019-2020). *DeepSpeech*. Adferwyd 08 2020 o GitHub: https://github.com/mozilla/DeepSpeech

Mozilla. (2020). *Common Voice Cymraeg*. Adferwyd 07 2020 o Common Voice: https://voice.mozilla.org/cy/speak

Mulvenna, M., Hutton, A., Coates, V., Martin, S., Todd, S., Bond, R., & Moorhead, A. (2017, 01 24). Views of Caregivers on the Ethics of Assistive Technology Used for Home Surveillance of People Living with Dementia. *Neuroethics, 10*, 255-266.

Najafian, M., Safavi, S., Hanani, A., & Russell, M. (2014). Acoustic model selection using limited data for accent robust speech recognition. *European Signal Processing Conference* (tt. 1786-1790). Lisbon: IEEE.

Neri, A., Cucchiarini, C., & Strik, W. (2003). Automatic Speech Recognition for second language learning: How and why it actually works. *15th ICPhS* (tt. 1157-1160). Barcelona: ISBN.

Niehues, J., Ha, T.-l., Peter, J.-T., Sennrich, R., Williams, P., Frank, S., . . . Deksne, D. (2018, 01 26). *Final Report: Morphologically Rich languages.* Adferwyd 09 2020 o QT21: https://www.google.com/search?client=safari&rls=en&q=final+report:+morphologically+rich+languages+qt21&ie=UTF-8&oe=UTF-8

Nordquist, R. (2020, 08 26). *Definition and Examples of Language Varieties.* Adferwyd 09 2020 o ThoughtCo: https://www.thoughtco.com/language-variety-sociolinguistics-1691100

Office for National Statistics. (2013). *Office for National Statistics.* Adferwyd 04 14, 2021 o Language in England and Wales: 2011: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language/articles/languageinenglandandwales/2013-03-04

Oftedal, M. (1969). A New Approach to North Welsh Vowels. *Lochlann, 4*, 243-269.

Oftedal, M. (1985). *Lenition in Celtic and Insular Spanish.* Oslo.

Ortega, L. (2009). *Understanding Second Language Acquisition.* Llundain: Routledge.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). *Librispeech: an ASR Corpus Based on Public Domain Audio Books.* Adferwyd 06 2020 o danielpovey.com: http://www.danielpovey.com/files/2015_icassp_librispeech.pdf

Pandey, A., Kumar, R., Nellore, B. T., Teja, K. S., & Gangashetty, S. V. (2018). Phonetically Balanced Code-Mixed Speech Corpus for Hindi-English Automatic Speech Recognition. *LREC 2018*, (tt. 1480-1484).

Pellegrini, T., & Lamel, L. (2009). Automatic word decompounding for ASR in a morphologically rich language: application in Amharic. *IEEE Transactions on Audio Speech and Language Processing. 17*, tt. 863-873. IEEE.

Pierce, J. R. (1969). Whither Speech Recognition? *The Journal of the Acoustical Society of America, 46*, 1049-1051.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., . . . Veselý, K. (2011). The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (tt. 1-4). Waikoloa: ASRU.

Prabhavalkar, R., Rao, K., Sainath, T. N., Li, B., Johnson, L., & Jaitly, N. (2017). A Comparison of Sequence-to-Sequence Models for Speech Recognition. *INTERSPEECH 2017* (tt. 939-943). Sweden: ISCA 2017.

Prys, D. (2006). Setting the Standards: ten years of Welsh terminology work. Yn P. t. Hacken, *Terminology, Computing and Translation.* Gunnter Narr Verlag, 41-55: Tübingen.

Prys, D. (2008). The development and acceptance of electronic resources for Welsh. Yn V. Lyding (Gol.), *Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics (LULCL II)* (tt. 33-41). Bozen: EURAC.

Prys, D., & Jones, D. B. (2015). National Language Technologies Portals for LRLs: A Case Study. *Human Language Technology: Challenges for Computer Science and Linguistics*, 420-429.

Prys, D., & Jones, D. B. (2018). Gathering Data for Speech Technology in the Welsh Language. Yn C. Soria, L. Besacier, & L. Pretorius (Gol.), *Proceedings of LREC 2018 Workshop "CCURL 2018 - Sustaining Knowledge Diversity in the Digital Age"*, (tt. 56-61).

Prys, D., Jones, D. B., & Prys, G. (2020). Coherent Planning for Language Technology Developments and Language Revitalization. Yn G. A. al. (Gol.), *LT4All proceedings* (tt. 367-370). Paris: ELRA.

Prys, D., Prys, G., & Jones, D. B. (2016). Cysill Ar-lein: A Corpus of Written Contemporary Welsh Compiled from an On-line Spelling and Grammar Checker. *Proceedings of the Language Resources and Evaluation Conference (LREC)* (tt. 3261-3264). Portorož: The European Language Resources Association.

Prys, D., Williams, B., Hicks, B., Jones, D., Chasaide, A. N., Gobl, C., . . . Dhonnachadha, E. U. (2004). WISPR: Speech Processing Resources for Welsh and Irish. *SALTMIL Workshop at LREC 2004: First Steps for Language Documentation of Minority Languages* (tt. 68-71). Lisbon: ELRA.

Prys, M. (2016, 01). *Style in the vernacular and on the radio: code-switching and mutation as stylistic and social markers in Welsh.* Adferwyd 08 2020 o Prifysgol Bangor: https://research.bangor.ac.uk/portal/en/theses/style-in-the-vernacular-and-on-the-radio-codeswitching-and-mutation-as-stylistic-and-social-markers-in-welsh(952270df-9881-48b5-9193-85b0b8c4eecd).html

Prys, R. (2019, 08 02). *Common Voice Cymraeg yn yr Eisteddfod.* Adferwyd 07 2020 o Meddal: Meddalwedd Cymraeg: https://www.meddal.com/meddal/?p=1753

Prys, R. (2019, 06 08). *Dathlu blwyddyn o Common Voice Cymraeg.* Adferwyd 07 2020 o Meddal: Meddalwedd Cymraeg: https://www.meddal.com/meddal/?p=1718

Prys, R. (2019, 06 28). *Y Cynulliad Cenedlaethol yn Cefnogi Common Voice Cymraeg.* Adferwyd 07 2020 o Meddal: Meddalwedd Cymraeg: https://www.meddal.com/meddal/?p=1744

Rabiner, L. R. (1984, 01). The acoustics, speech, and signal processing society - A historical perspective. *IEEE ASSP Magazine, 1*(1), 4-10.

Rees, I. W. (2015). Phonological variation in Mid Wales. *Studia Celtica, 49*(1), 149-174.

Rees, I. W. (2016). "Dim Sôn am Dduw na Dyn": Ar Drywydd yr 'U Ogleddol' yng Nghanolbarth Cymru. *Gwerddon*(22), 47-74.

Rees, I. W., & Morris, J. (2018). Astudiaeth o ganfyddiadau tiwtoriaid Cymraeg i Oedolion o anawsterau ynganu ymhlith dysgwyr yr iaith. *Gwerddon, 27*, 39-66.

Ritchie, W. C., & Bhatia, T. K. (2012). Social and Psychological Factors in Language Mixing. Yn T. K. Bhatia, & W. C. Ritchie, *Handbook of Bilingualism and Multilingualism* (tt. 275-390). Wiley.

Sak, H., Saraşlar, M., & Güngör, T. (2009). Integrating morphology into automatic speech recognition. *2009 IEEE Workshop on Automatic Speech Recognition & Understanding* (tt. 354-358). Merano: IEEE.

Sak, H., Saraçlar, M., & Güngör, T. (2010). Morphology-based and sub-word language modeling for Turkish speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010* (tt. 5402-5405). Dallas: ISCA.

Samuel, A. L. (1959, 07). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal*, 210-229.

Schultz, T., & Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication, 35*, 31-51.

Selinker, L. (1972). Interlanguage. *IRAL-International review of Applied Linguistics in Language Teaching, 10*(3), 209-232.

Shahin, M., Ahmed, B., McKechnie, J., Ballard, K., & Guitierrez-Osuna, R. (2014). A Comparison of GMM-HMM and DNN-HMM Based Pronunciation Verification Techniques for Use in the Assessment of Childhood Apraxia of Speech. *INTERSPEECH 2014* (tt. 1583-1587). Singapore: ISCA.

Spasić, I., Owen, D., Knight, D., Artemiou, A., Batchelor, C., Bell, E., . . . Raghal, Ó. (2019). Proceedings of the Celtic Language Technology Workshop. *Dublin Machine Translation Summit 2019.* Dublin.

Sri, K. V., Srinivasan, M., Nair, R. R., Priya, K. J., & Gupta, D. (2020). Kaldi recipe in Hindi for word level recognition and phoneme level transcription. *Procedia Computer Science*(171), 2476-2485.

Sriram, A., Jun, H., Satheesh, S., & Coates, A. (2017, 08 21). *Cold Fusion: Training Seq2Seq Models Together with Language Models.* Adferwyd 07 2020 o arXiv: https://arxiv.org/abs/1708.06426

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Shlüter, P., Przybyszewski, M., & Gilbro, S. (2014, 08 29). An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation, 48*, 679-707.

Strik, H., Neri, A., & Cucchiarini, C. (2008). Speech technology for language tutoring. *Proceedings of LangTech*, (tt. 1-4).

Tachbelie, M. Y., Abate, S. T., & Besacier, L. (2014, 01). Using different acoustic, lexical and language modeling units for ASR of an under-resourced language - Amharic. *Speech Communication, 54*, 181-194.

Tarján, B., Szaszák, G., Fegyó, T., & Mihajlik, P. (2019). Investigation on n-gram approximated RNNLMs for recognition of morphologically rich speech. Yn C. Martín-Vide, M. Purver, & S. Pollak (Gol.), *7th International Conference, SLSP 2019* (tt. 223-234). Ljubljana: Springer International Publishing.

Techiaith. (2018). *Lleisiwr.* Adferwyd 08 2020 o Techiaith: https://lleisiwr.techiaith.cymru

Techiaith. (2019). *Macsen.* Adferwyd 07 2020 o Porth Technolegau Iaith Cenedlaethol Cymru: http://techiaith.cymru/pecynnau/macsen/

Templeton, G. (2020, 09 15). *Voice Assistants Compared: Comparing the Language Support of Siri, Alexa, and Google Assistant.* Adferwyd 09 2020 o Globalme: https://www.globalme.net/blog/language-support-voice-assistants-compared/

Thomas, B., & Thomas, P. W. (1989). *Cymraeg, Cymrâg, Cymrêg... Cyflwyno'r Tafodieithoedd.* Caerdydd: Gwasg Taf.

Thomas, E. M., & Webb-Davies, P. (2017, 08 04). *Agweddau ar Ddwyieithrwydd.* Adferwyd 06 2020 o Porth y Coleg Cymraeg Cenedlaethol: https://llyfrgell.porth.ac.uk/View.aspx?id=3001~4e~cunoKUqT

Thomas, P. W. (1996). *Gramadeg y Gymraeg.* Caerdydd: Gwasg Prifysgol Cymru.

Thorne, D. A. (1996). *Gramadeg Cymraeg.* Llandysul: Gwasg Gomer.

Tjalve, M., & Huckvale, M. (2005). Pronunciation variation modelling using accent features. *INTERSPEECH 2005 - EUROSPEECH* (tt. 1341-1344). Lisbon: ISCA.

Tomlin, R. S. (1986). *Basic word order: Functional principles.* Llundain: Croom Helm.

Toshinwal, S., Sainach, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., & Rao, K. (2018). Multilingual Speech Recognition with a Single End-to-End Model. *IEEE 2018 International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (tt. 4899-4903). Calgary: IEEE.

Upadhyaya, P., Farooq, O., Abidi, M. R., & Varshney, Y. V. (2017). Continuous Hindi Speech Recognition Model Based on Kaldi ASR Toolkit. *UEEE WiSPNET*, (tt. 786-789).

Uraga, E., & Gamboa, C. (2004). VOXMEX Speech Database: Design of a Phonetically Balanced Corpus. *LREC 2004*, (tt. 1471-1474).

Wang, D., Wang, X., & Lv, S. (2019, 08 07). An Overview of End-to-End Automatic Speech Recognition. *Symmetry, 11*, 1018-1045.

Watanabe, S., Hori, T., & Hershey, J. (2017). Language Independent End-to-End Architecture for Joint Language and Speech Recognition. *IEEE Workshop on Automatic Speech Recognition and Understanding* (tt. 1-9). IEEE.

Watkins, A. (1961). *Ieithyddiaeth: Agweddau ar astudio iaith.* Caerdydd: Gwasg Prifysgol Cymru.

Watkins, T. A. (2002). Welsh. Yn M. J. Ball, & J. Fife, *The Celtic Languages* (tt. 289-348). Llundain: Routledge.

Webb-Davies, P., & Shank, C. (2020, 03). Defnydd hanesyddol a chyfoes o 'mynd i' yn y Gymraeg: Astudiaeth o ramadegoli fel newid ieithyddol. *Gwerddon, 30*, 23-39.

Williams, B. (1994). Diphone Synthesis for the Welsh Language. *International Conference on Spoken Language Processing (ICSLP)* (tt. 739-742). Yokohama: ISCA.

Williams, B. (1994). Welsh letter-to-sound rules: rewrite rules and two-level rules compared. *Computer Speech and Language, 8*, 261-277.

Williams, B. (1995). Text-to-speech Synthesis for Welsh and Welsh English. *EUROSPEECH* (tt. 1113-1117). Madrid: ISCA.

Williams, B. (1999). A Welsh Speech Database: Preliminary Results. *Proceedings of Eurospeech* (tt. 2283-2286). Budapest: ISCA.

Williams, B., Jones, R. J., & Uemlianin, I. (2006). Tools and resources for speech synthesis arising from a Welsh TTS project. *Language Resources Evaluation Conference (LREC)* (tt. 2574-2577). Genoa: ELRA.

Williams, S. J. (1980). *Elfennau Gramadeg Cymraeg.* Caerdydd: Gwasg Prifysgol Cymru.

Willis, D. (2019). Dialect syntax as a testbed for models of innovation and change: Modals and negative concord in the Syntactic Atlas of Welsh Dialects. *Syntactic Atlas of Welsh Dialects. Glossa: a journal of general linguistics, 4*(1), 31: 1-30.

WISPR. (diddyddiad). *WISPR*. Adferwyd 08 2020 o WISPR: http://www.e-gymraeg.org/wispr/index_en.htm

Wooldridge, D. (2011). *Gwella Cysill at ddefnydd Cyfieithwyr: adnabod ymyrraeth gan yr iaith Saesneg mewn testunau Cymraeg.* Adferwyd 07 2020 o Cymdeithas Cyfieithwyr Cymru: https://www.cyfieithwyr.cymru/files/THESIS_DAWN_WOOLDRIDGE1.pdf

Wright, R. (2006). Intra-speaker variation and units in human speech perception and ASR. *Proceedings of the Speech Recognition and Intrinsic Variation Workshop* (tt. 39-42). Tolouse: ISCA.

Wu, Y., Zhang, R., & Rudnicky, A. (2007). Data selection for speech recognition. *IEEE workshop on automatic speech recognition and understanding* (tt. 562-565). Pittsburgh: IEEE.

Yılmaz, E., Heuvel, H. v., & Leeuwen, D. v. (2016). "Investigating Bilingual Deep Neural Networks for Automatic Recognition of Code-switching Frisian Speech", 5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia. *Procedia Computer Science, 81*, 159-166.

Yılmaz, E., Heuvel, H. v., Dijkstra, J., Velde, H. V., Kampstra, F., Algra, J., & Leeuwen, D. V. (2016). Open Source Speech and Language Resources for Frisian. *INTERSPEECH. 81*, tt. 1536-1540. San Francisco: ISCA.

Yarowsky, D., Ngai, G., & Wicentowsky, R. (2001). Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. *Proceedings of the First International Conference on Human Language Technology Research*, (tt. 1-8).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. (., . . . Zhang, C. (2015, 12). *The HTK Book version 3.5a.* Adferwyd 06 2020 o HTK3: http://htk.eng.cam.ac.uk/docs/docs.shtml

Yu, D., & Deng, L. (2015). *Automatic Speech Recognition: A Deep Learning Approach.* Llundain: Springer.

Zhang, C., & Hansen, J. H. (2019). Advancements in whispered speech detection for interactive/speech systems. Yn H. A. Patil, A. Neustein, & M. Kulshreshtha, *Signal and Acoustic Modeling for Speech and Communication Disorders* (tt. 9-32). De Gruyter.

Rhestr Atodiadau

Atodiad A. Promptiau Paldauro 1.0, 2.0 a 3.0

```
*/sample1 LLEUAD MELYN AELODAU SIARAD FFORDD YMLAEN CEFNOGAETH HELEN
*/sample2 GWRAIG OREN DIWRNOD GWAITH MEWN EISTEDDFOD DISGOWNT IDDO
*/sample3 OHERWYDD ELLIW AWDURDOD BLYNYDDOEDD GWLAD TYWYSOG LLYW UWCH
*/sample4 RHYBUDDIO ELEN UWCHRADDIO HWNNW BEIC CYMRU RHOI AELOD
*/sample5 RHAI STEROID CEFNOGAETH FELEN CAU GAREJ ANGAU YMHLITH
*/sample6 GWNEUD IAWN UN DWEUD LLAIS WEDI GYDA LLYN
*/sample7 LLIW YNG_NGHYMRU GWNEUD ROWND YCHYDIG WY YN LLAES
*/sample8 HYN NEWYDDION AR ROEDD PAN LLUN MELIN SYCHU
*/sample9 YCHYDIG GLIN WRTH HUW AT NHW BOD BYDD
*/sample10 YN_UN ER_MWYN NEU_DDYSGU HYD_YN_OED TAN OND_FE_AETH ATI
*/sample11 Y_GYMDEITHAS YNO_YN_FUAN MAWR GANRIF AMSER DECHRAU CYFARFOD
*/sample12 PRIF RHAID_BOD RHEINI SADWRN SY'N_COFIO CYNTAF RHAID_CAEL
*/sample13 DROS_Y_FFORDD GWASANAETH BYDDAI'R_RHESTR HYD LLYGAID LLOEGR
*/sample14 CEFN TEULU ENWEDIG OND_MAE Y_TU Y_PRYD DI-HID PETH HEFYD
*/sample15 MORGAN ETO YMA DDEFNYDDIO BACH YN_WIR DIWEDD LLENYDDIAETH
*/sample16 YM_MRYSTE NATUR OCHR MAE_HI NEWID DY_GYMORTH NES GWAHANOL
*/sample17 I_DDOD CYNGOR ATHRAWON BYCHAN NEU DIGWYDD HUD MYND_I_WELD
*/sample18 EI_GILYDD CYFFREDIN HUNAIN LLE CYMDEITHASOL Y_LLE UNWAITH
*/sample19 I_TI NEWYDD YSGRIFENNU Y_GWAITH DARLLEN FYDDAI ADDYSG DAETH
*/sample20 LLYWODRAETH OND HYNNY ESGOB CYRRAEDD A_BOD GWRS CEIR
*/sample21 RHAID_GWELD CHWARAE NAD_OEDD WEDYN FLWYDDYN OND_NID ARDAL
*/sample22 BUASAI HANES DDIWEDDAR WEDI_CAEL O_BOBL MERCHED FFILM CAFODD
*/sample23 AWDUR NA OEDD_MODD DOD YR_HEN GEN_I OLAF DDECHRAU
*/sample24 DYNA DDIGON I_BEIDIO BYNNAG RHAN TRWY AM_Y_LLYFR Y_CYFNOD
*/sample25 ATHRO ANIFEILIAID POB O_FEWN YN_GWNEUD CARTREF ELFENNAU
*/sample26 ER_ENGHRAIFFT BRON YN_FWY AR_GAEL SYLW EDRYCH_ARNO ARALL
*/sample27 CYHOEDDUS UN_PRYD CLYWED OHONOM EI_FOD AROS GWYRDD_GOLAU
*/sample28 YN_EI_GWEN MAI DOD_O_GYMRU PERSONOL ALLAN WRTH_Y_FFENESTR
*/sample29 YSTYR DDA ARBENNIG MAE'N_BWYSIG OEDDWN FARW NIFER_O_WYAU MAER
*/sample30 AMERICA AR_GYFER IAITH BELLACH GENEDLAETHOL ATEB AT_Y_BONT
*/sample31 AR_Y_CEFN AC_ROEDD NESAF I_GYD DOEDD_DIM CYNNWYS AMLWG
*/sample32 AMGYLCHIADAU GWEITHWYR FY_MAM AC_YN_LLOGI PETHAU UNRHYW DRWS
*/sample33 EVANS YN_MYND CORFF NEB EGLWYS CAFWYD SEF AR_EI
*/sample34 DATBLYGU AC_ATI TRADDODIAD YN_BYW OND_HEFYD Y_DYDD WILLIAMS
*/sample35 DOSBARTH YR_UN FOD_YN_FAWR NI YR_YSGOL AIL_GANRIF AM NID
*/sample36 GOFYNNODD GWYBOD LLAWER RHYWBETH O_RYWLE CHWILIO_AM ODDI_AR
*/sample37 CYNLLUN CYCHWYN DIOLCH LLYFR YN_Y_BLAEN DAN I_DDIM CYN
*/sample38 I'R_DDE DDYLETSWYDD HI MAE'N_HWYR DROS MEGIS MILLTIR ADEG
*/sample39 AMBELL YR_OGOF YNA LERPWL YSGOLION PARC DAL PLANT
*/sample40 MAM OEDD_HWN IFANC GELLIR OESOEDD_CANOL CAPEL YSGOL MLYNEDD
*/sample41 O_GWMPAS HON WEITHIAU ERBYN_HYN STORI I_FOD GANDDO YN_CAEL
*/sample42 SIR_BENFRO GWELD GILYDD OND_DOEDD OES UN_O'CH_FFRINDIAU YSTOD
*/sample43 DDIM OND_PAN EDRYCH WRTH_GWRS A_PHAN YSTYRIED WEDI_BOD
```

Atodiad B. Promptiau Paldaruo 4.0

```
*/sample1 LLEUAD MELYN AELODAU SIARAD FFORDD YMLAEN CEFNOGAETH HELEN
*/sample2 GWRAIG OREN DIWRNOD GWAITH MEWN EISTEDDFOD DISGOWNT IDDO
*/sample3 OHERWYDD ELLIW AWDURDOD BLYNYDDOEDD GWLAD TYWYSOG LLYW UWCH
*/sample4 RHYBUDDIO ELEN UWCHRADDIO HWNNW BEIC CYMRU RHOI AELOD
*/sample5 RHAI STEROID CEFNOGAETH FELEN CAU GAREJ ANGAU YMHLITH
*/sample6 GWNEUD IAWN UN DWEUD LLAIS WEDI GYDA LLYN
*/sample7 LLIW YNG_NGHYMRU GWNEUD ROWND YCHYDIG WY YN LLAES
*/sample8 HYN NEWYDDION AR ROEDD PAN LLUN MELIN SYCHU
*/sample9 YCHYDIG GLIN WRTH HUW AT NHW BOD BYDD
*/sample10 YN_UN ER_MWYN NEU_DDYSGU HYD_YN_OED TAN OND_FE_AETH ATI
*/sample11 Y_GYMDEITHAS YNO_YN_FUAN MAWR GANRIF AMSER DECHRAU CYFARFOD
*/sample12 PRIF RHAID_BOD RHEINI SADWRN SY'N_COFIO CYNTAF RHAID_CAEL
*/sample13 DROS_Y_FFORDD GWASANAETH BYDDAI'R_RHESTR HYD LLYGAID LLOEGR
*/sample14 CEFN TEULU ENWEDIG OND_MAE Y_TU Y_PRYD DI-HID PETH HEFYD
*/sample15 MORGAN ETO YMA DDEFNYDDIO BACH YN_WIR DIWEDD LLENYDDIAETH
*/sample16 YM_MRYSTE NATUR OCHR MAE_HI NEWID DY_GYMORTH NES GWAHANOL
*/sample17 I_DDOD CYNGOR ATHRAWON BYCHAN NEU DIGWYDD HUD MYND_I_WELD
*/sample18 EI_GILYDD CYFFREDIN HUNAIN LLE CYMDEITHASOL Y_LLE UNWAITH
*/sample19 I_TI NEWYDD YSGRIFENNU Y_GWAITH DARLLEN FYDDAI ADDYSG DAETH
*/sample20 LLYWODRAETH OND HYNNY ESGOB CYRRAEDD A_BOD GWRS CEIR
*/sample21 RHAID_GWELD CHWARAE NAD_OEDD WEDYN FLWYDDYN OND_NID ARDAL
*/sample22 BUASAI HANES DDIWEDDAR WEDI_CAEL O_BOBL MERCHED FFILM CAFODD
*/sample23 AWDUR NA OEDD_MODD DOD YR_HEN GEN_I OLAF DDECHRAU
*/sample24 DYNA DDIGON I_BEIDIO BYNNAG RHAN TRWY AM_Y_LLYFR Y_CYFNOD
*/sample25 ATHRO ANIFEILIAID POB O_FEWN YN_GWNEUD CARTREF ELFENNAU
*/sample26 ER_ENGHRAIFFT BRON YN_FWY AR_GAEL SYLW EDRYCH_ARNO ARALL
*/sample27 CYHOEDDUS UN_PRYD CLYWED OHONOM EI_FOD AROS GWYRDD_GOLAU
*/sample28 YN_EI_GWEN MAI DOD_O_GYMRU PERSONOL ALLAN WRTH_Y_FFENESTR
*/sample29 YSTYR DDA ARBENNIG MAE'N_BWYSIG OEDDWN FARW NIFER_O_WYAU MAER
*/sample30 AMERICA AR_GYFER IAITH BELLACH GENEDLAETHOL ATEB AT_Y_BONT
*/sample31 AR_Y_CEFN AC_ROEDD NESAF I_GYD DOEDD_DIM CYNNWYS AMLWG
*/sample32 AMGYLCHIADAU GWEITHWYR FY_MAM AC_YN_LLOGI PETHAU UNRHYW DRWS
*/sample33 EVANS YN_MYND CORFF NEB EGLWYS CAFWYD SEF AR_EI
*/sample34 DATBLYGU AC_ATI TRADDODIAD YN_BYW OND_HEFYD Y_DYDD WILLIAMS
*/sample35 DOSBARTH YR_UN FOD_YN_FAWR NI YR_YSGOL AIL_GANRIF AM NID
*/sample36 GOFYNNODD GWYBOD LLAWER RHYWBETH O_RYWLE CHWILIO_AM ODDI_AR
*/sample37 CYNLLUN CYCHWYN DIOLCH LLYFR YN_Y_BLAEN DAN I_DDIM CYN
*/sample38 I'R_DDE DDYLETSWYDD HI MAE'N_HWYR DROS MEGIS MILLTIR ADEG
*/sample39 AMBELL YR_OGOF YNA LERPWL YSGOLION PARC DAL PLANT
*/sample40 MAM OEDD_HWN IFANC GELLIR OESOEDD_CANOL CAPEL YSGOL MLYNEDD
*/sample41 O_GWMPAS HON WEITHIAU ERBYN_HYN STORI I_FOD GANDDO YN_CAEL
*/sample42 SIR_BENFRO GWELD GILYDD OND_DOEDD OES UN_O'CH_FFRINDIAU YSTOD
*/sample43 DDIM OND_PAN EDRYCH WRTH_GWRS A_PHAN YSTYRIED WEDI_BOD
*/sample44 RHAWN UNGELLOG CHWITFFORDD DEHEUBARTH ROBERTS THAW HAWYS DDUW
*/sample45 TWYMYN DEIO STEPDIR EINGION DUC GOETS ABERDAUGLEDDAU
LLETCHWITH
*/sample46 PAILL GAWSAI GYW ACHUB PAWB COIL MAENGWYN FRIWYDD
*/sample47 HEI OSGOI BLAENHONDDAN GAEAFGWSG DOE DREIER LLANGURIG PITSA
*/sample48 UWCHBEN IACHAWDWRIAETH LLUWCHWYNT NGHYN JOBS CLYWCH ARFBAIS
TEWDWR
*/sample49 WYAU LLANDDOWROR DULL BOEN LLWYNHENDY MHOWYS TREFDRAETH
CADMIWM
*/sample50 BEUNO TROTSCÏAETH PEIPEN CERRIG_LLWYDION FFAWT TANGNEFEDD
FWYELL HOE
*/sample51 BWYTÄWR GWEU NGHAWL RHOSHIRWAUN HYWYN ANHUNEDD PWLLHELI
ANGHOFUS
*/sample52 CWMLLAN UTHR RHYWBETH PAENT GULDDAIL MEWN ANGHEUOL RHAI
*/sample53 BWFFE IDDEW DAUFINIOG GORUWCHNATURIOL DYWYLL ENGHREIFFTIOL
CHOED HOPCYN
```

*/sample54 GWIBDAITH PIBGOD UWD GAREJ LLYWYDD HOYW TROWSUS HYLL
*/sample55 TWNGSTEN AIFFT ATHLETAU FFOWC DDWYIEITHOG CARNHEDRYN LLOI
BAWB
*/sample56 NANTCLWYD ACHAU CULHWCH EWTHANASIA MAIP WDDF CIWB GWAYWFFON
*/sample57 WYCH JAMAICA CAU PUPUR CROESHOELIO MENYW GWRYWDOD BUWCH
*/sample58 FRITHGRAIG CERNYWIAID GEUFFORDD PONTSENNI CASTELLNEWYDD LOWRI
MYW FROWNGOCH
*/sample59 PROSSER LLAETH LLAW CEGDDU STOW TROELL TEITL MOI
*/sample60 WTHIO BEIAU WRTHBLAID HOW NOETHNI DEWCH PROJECT AUR
*/sample61 BLODEUYN MHARIS BOWNSIO DDAETH CNICHT ALPAU THYWYN CHYFF
*/sample62 HUFEN CELLBILEN ÛYDD FOWLER PHWYSAU NANTPERIS BODFFORDD
CYWYDD
*/sample63 LECWYDD RWSIAIDD UCHAFSWM FFIWSIA MHEN MÊTS LLEILL LLEOEDD
*/sample64 LLEUCU FELINHELI GWRHYD LLYWN NAWDDSANT RHEIBIWR PWLLMEURIG
HEWL
*/sample65 DAUWYNEBOG CELLRANIAD RUFFYDD LLANUWCHLLYN DWYFFURF CEFNGRWM
RHOI EFENGYL
*/sample66 DDEUFAEN NHW FFOTOGRAFFAU MHORTIWGAL RHITHDYB ACSIWN IAWN
LLOER
*/sample67 PORTHCAWL TWPSYN NANTLLE TABLOID NHYWYN AMHLEIDIOL DDOI
DDREWLLYD
*/sample68 JYNGL TÖWR COEDPOETH PENRHIWLLAN BRIWFWYD DOI IEUAF BANJO
*/sample69 FODCA TREFHEDYN BYWGRAFFIAD COFIWCH BAEDD BENTHYG ATSAIN
GRIFFITHS
*/sample70 MHIC LLONGAU CHWARAEWR FUWCH THAI PONTSTICILL FEURIG DREWGI
*/sample71 TEITHIWR BAICH FEWN HUWS PNAWN RHYTHM FAWR GRONGAER
*/sample72 ALLGO CLAWDDNEWYDD CYW CHAEAU SUO SAETS COWBOIS GOCHDDU
*/sample73 CANHWYLLAU BOLSIEFIC NGHLWYD LLWYNGWAIR TEIM TAEOG ALLDAFLIAD
MUHAMMAD
*/sample74 ATALNWYD EFFAITH EICH GLYWSOCH HWYRHAU CALSIWM DEUGAIN
SIOEFERCH
*/sample75 CAERDROEA FYW NGHWM CWMCERWYN RHWNG CAWG GWAWR PWYTH
*/sample76 DUWYNT MOREAU RHYWFAINT DDWFN RHYW POWDWR SIOEAU LOEGR
*/sample77 THEULUOEDD TOES PORTHAETHWY CIBWTS RHAEADR LLIW MINOAIDD
NYMFF
*/sample78 MAGDALEN CEWRI FFEUEN CLWYFAU PUW SIPSIWN LLAI FRONHAUL
*/sample79 SOIA DEUAWD PRAWF ROIS TEULU BYW DDAW AMHEUS
*/sample80 BWDHAETH BOTSWANA GEWYN HEDDIW EBBW STORÏWR LLANGEITHO
LLEISIAU
*/sample81 CAERHUN LLEW ARLLWYSIAD IEITHOEDD EHANGDIR CEULAN BONTDDU
NHRWYN
*/sample82 DDOE SECCO HIRHOEDLOG TYWYLL FYWYD CARNGUWCH BARHAUS HAUL
*/sample83 CAIO PIWS CORHELGI MAESLLYN FOICOTIO GIW JIN DAWCH
*/sample84 PWLLCROCHAN CAMFA RHEW CROESRYW CAIB FFACBYS CLUSTDLWS HIWMOR
*/sample85 GOWT JIWDO NIWCLEWS CORSIOG MEUDWYOL EINIOES BWÂU

Atodiad C. Sampl gynrychiadol o *Common Voice-cy* (13/02/19)

manylion lawrlwytho feriswn 13/02/19:

```
https://voice-prod-bundler-
ee1969a6ce8178826482b88e843c335139bd3fb4.s3.amazonaws.com/cv-corpus-
2019-02-13/

COMMONVOICE_DOWNLOAD_URL = 'https://voice-prod-bundler-
ee1969a6ce8178826482b88e843c335139bd3fb4.s3.amazonaws.com/cv-corpus-
3/cy.tar.gz'
```

dolen lawrlwytho'r corpws mwyaf diweddar (22/06/20):

https://voice.mozilla.org/cy/datasets

```
*/sample1 LLEUAD MELYN AELODAU SIARAD FFORDD YMLAEN CEFNOGAETH HELEN
*/sample2 GWRAIG OREN DIWRNOD GWAITH MEWN EISTEDDFOD DISGOWNT IDDO
*/sample3 OHERWYDD ELLIW AWDURDOD BLYNYDDOEDD GWLAD TYWYSOG LLYW UWCH
*/sample4 RHYBUDDIO ELEN UWCHRADDIO HWNNW BEIC CYMRU RHOI AELOD
*/sample5 RHAI STEROID CEFNOGAETH FELEN CAU GAREJ ANGAU YMHLITH
*/sample6 GWNEUD IAWN UN DWEUD LLAIS WEDI GYDA LLYN
*/sample7 LLIW YNG_NGHYMRU GWNEUD ROWND YCHYDIG WY YN LLAES
*/sample8 HYN NEWYDDION AR ROEDD PAN LLUN MELIN SYCHU
*/sample9 YCHYDIG GLIN WRTH HUW AT NHW BOD BYDD
*/sample10 YN_UN ER_MWYN NEU_DDYSGU HYD_YN_OED TAN OND_FE_AETH ATI

*/sample75 CAERDROEA FYW NGHWM CWMCERWYN RHWNG CAWG GWAWR PWYTH
*/sample76 DUWYNT MOREAU RHYWFAINT DDWFN RHYW POWDWR SIOEAU LOEGR
*/sample77 THEULUOEDD TOES PORTHAETHWY CIBWTS RHAEADR LLIW MINOAIDD
NYMFF
*/sample78 MAGDALEN CEWRI FFEUEN CLWYFAU PUW SIPSIWN LLAI FRONHAUL
*/sample79 SOIA DEUAWD PRAWF ROIS TEULU BYW DDAW AMHEUS
*/sample80 BWDHAETH BOTSWANA GEWYN HEDDIW EBBW STORÏWR LLANGEITHO
LLEISIAU
*/sample81 CAERHUN LLEW ARLLWYSIAD IEITHOEDD EHANGDIR CEULAN BONTDDU
NHRWYN
*/sample82 DDOE SECCO HIRHOEDLOG TYWYLL FYWYD CARNGUWCH BARHAUS HAUL
*/sample83 CAIO PIWS CORHELGI MAESLLYN FOICOTIO GIW JIN DAWCH
*/sample84 PWLLCROCHAN CAMFA RHEW CROESRYW CAIB FFACBYS CLUSTDLWS HIWMOR
*/sample85 GOWT JIWDO NIWCLEWS CORSIOG MEUDWYOL EINIOES BWÂU

(brawddegau UTI)
*/d042bcafd3cfc02184eeb4e3792a30c4 AC MAE GANDDI DRAWSTORIAD O BOBOL YN
GWEITHIO IDDI
*/1b9680d024802138bef4d7a6c0053d21 AC MAE HYN OLL CYN I BOBOL DDECHRAU
CANOLBWYNTIO AR YR ETHOLIAD O DDIFRIF
*/2ed1e27323ff721fe85525e2c6a5887e A DOEDD Y CANIATÂD HWN DDIM YN CAEL
EI ROI YN DDI-GWESTIWN
*/611145e341c75c25c447225c08cb9d3f AETH HWNNA MOR GLOU AC MOR ARAF AR YR
UN PRYD
*/3b473df687f4835f3bc4d5be4f6c3840 ALED ROBERTS YN DECHRAU AR Y GWAITH
AR UNWAITH
*/4574d86b5fcddaf2b19e4823f92745ec ALLAN NHW EISTEDD ROWND Y BWRDD A
CHAEL PRYD IAWN
*/69fd605291209821e7467f56f4583dbf ALLWN NI DDIM EU TRIN FEL UNRHYW
BLAID ARALL
*/a1dd4aa0bf84ebaa073344b7d709fc34 ANODD YW DOD O HYD I UNRHYW BETH DA
I'W DDWEUD AM FIS CHWEFROR
*/c2363ace81d7a608da078349d70826ce ATEBWYD FY NGHRI GAN YR EISTEDDFOD AR
FFURF Y BARIAU AR Y MAES
```

```
*/a579182d019800ac5e8d0a552594f334 BARN TRI AR DDIWRNOD I ANNOG POBL I
GOFRESTRU I BLEIDLEISIO
*/d07170b37f037d7d8d883a8ce7ba6c4f BELLACH HEFYD ROEDD POBL YN GALW
HEIBIO EIN CARTREF YN SIOP Y PENTRE

(promptiau Macsen)
*/0955a2f1228561a2e87ecbe5b04120d5 BETH FYDD Y TYWYDD FORY
*/c9d8244ce45dfc242c50bf6a5032cdf0 BETH FYDD TYWYDD YFORY
*/6d2ce19e889fe0c3a7c6705cbeb95aab BETH OEDD NEWYDDION DDOE
*/f85141ae057cdb552250a5af581b73b8 BETH OEDD PENAWDAU DDOE
*/6e85420ad27aebfb0e2059451c1c1ffe BETH OEDD Y NEWYDDION DDOE
*/2d6312c5deef949bc0431ffdd86038a5 BETH OEDD Y PENAWDAU DDOE
*/bdd68790ad9de967d7c5fc43695a789b BETH YDY'R NEWYDDION
*/f8d207b75bca04145c974f1dcc94c9d4 BETH YDY'R PENAWDAU
*/a9f82f148af6a0e19794734f5fcede22 BETH YDY'R TYWYDD FORY
*/dd121c8c59260f5e2f9e3f1be62e83c9 BETH YDY'R TYWYDD HEDDIW
*/17ce693346b4c5fd93f738f4c0ab60c9 BETH YDY'R TYWYDD YFORY

(brawddegau Indeg)
*/01ab72b92f6829846eb58c2bbb538bca DYDW I DDIM YN BWRIADU BOD YNG
NGHAERDYDD DROS Y GWYLIAU
*/0a9463ca8f7e5414f674a35e9a50636a MAE ANGEN I TI OFYN AM BETH HOFFET TI
GAEL YN Y BWYTY
*/c59edf7c3bcd0f26134f56c19af0cc30 OEDD RHAID I TI DDWEUD NAD OEDDET
TI'N GWYBOD UNRHYW BETH
*/daf7ed5512b55176a4ae89d054e8d6de DOEDD DIM ANGEN DWEUD DIM YN Y DIWEDD
*/663134da017dac057478c51d12c13def MAE ANGEN TREFNU BWYD A DIOD AR GYFER
YR YMARFER AM SAITH
*/57ea23cc36adf9947caffd155620992a RWYT TI LAWER GWELL ERS YR ANNWYD
*/8f7137fd807b1a494695d6c0df653c6d OEDD ANGEN I CHI DDILYN CAMAU
ARBENNIG WRTH FYND AR YR ADEG YNA
*/1bebe9de8d7d34442e1c39f6c8e6c34f ROEDD HI'N GYFNOD ANODD GYDA'R GWYNT
GARW OEDD YN BRWYDRO GYDA'R COED
*/90368cc3ef2e9c15d682cb5ff2b6c135 MAE E'N DAL I FOD O GWMPAS YR ARDAL
*/7c07fa390a944d8157aa392f2a7d973b MAE ANGEN RHOI'R CYFAN I SYCHU O
FLAEN Y TÂN
*/2a15916c5d3ab2889fba9dce28d2bab9 YN LLE YDYCH CHI'N BYW
```

Atodiad D. Canlyniadau plygiadau Paldaruo 2.0 HTK a *Kaldi*

Canlyniadau plygiadau Paldaruo 2.0 HTK:

plyg	WER%
1	14.84
2	14.83
3	14.06
4	15.35
5	14.33
6	15.03
7	13.85
8	14.63
9	15.25
10	14.84
Cyfartaledd (WER%)	14.70
Gwyriad Safonol (N)	0.49
Cyfeiliornad Safonol (N)	0.16
Amrywiant	0.24

Canlyniadau plygiadau Paldaruo 2.0 *Kaldi* (tri3b):

plyg	WER% (tri3b)
1	6.59
2	3.61
3	4.54
4	5.07
5	4.42
6	3.22
7	3.56
8	3.79
9	4.61
10	4.11
Cyfartaledd (WER%)	4.38
Gwyriad Safonol (N)	1.02
Cyfeiliornad Safonol (N)	0.32
Amrywiant	1.05

Atodiad E. Canlyniadau plygiadau Paldaruo 3.0 HTK a *Kaldi*

Canlyniadau plygiadau Paldaruo 3.0 HTK:

plyg	WER%
1	14.84
2	14.83
3	14.06
4	15.35
5	14.36
6	15.03
7	13.85
8	14.63
9	15.25
10	14.84
Cyfartaledd (WER%)	14.70
Gwyriad Safonol (N)	0.49
Cyfeiliornad Safonol (N)	0.16
Amrywiant	0.24

Canlyniadau plygiadau Paldaruo 3.0 *Kaldi* (tri3b):

plyg	WER% (tri3b)
1	3.94
2	4.46
3	3.92
4	3.80
5	3.87
6	4.08
7	3.48
8	7.07
9	3.82
10	4.27
Cyfartaledd (WER%)	4.27
Gwyriad Safonol (N)	1.02
Cyfeiliornad Safonol (N)	0.32
Amrywiant	1.04

Atodiad F. Canlyniadau plygiadau Paldaruo 4.0 HTK a *Kaldi*

Canlyniadau plygiadau Paldaruo 4.0 HTK:

plyg	WER%
1	17.15
2	15.76
3	16.32
4	15.8
5	16.57
6	15.76
7	16.39
8	16.46
9	16.09
10	16.09
Cyfartaledd (WER%)	16.24
Gwyriad Safonol (N)	0.44
Cyfeiliornad Safonol (N)	0.14
Amrywiant	0.19

Canlyniadau plygiadau Paldaruo 4.0 *Kaldi* (tri3b):

plyg	WER% (tri3b)
1	4.96
2	4.73
3	4.17
4	4.77
5	4.54
6	7.00
7	4.41
8	4.04
9	4.74
10	3.86
Cyfartaledd (WER%)	4.96
Gwyriad Safonol (N)	4.73
Cyfeiliornad Safonol (N)	4.17
Amrywiant	4.77

Atodiad G. Canlyniadau plygiadau Common Voice-cy *Kaldi*

Canlyniadau plygiadau *Common Voice-cy Kaldi*:

plyg	WER%
1	3.56
2	3.11
3	3.56
4	3.18
5	3.52
6	3.50
7	3.36
8	3.47
9	3.53
10	3.65
Cyfartaledd (WER%)	3.44
Gwyriad Safonol (N)	0.17
Cyfeiliornad Safonol (N)	0.06
Amrywiant	0.03

Atodiad H. Sampl gynrychiadol o ddata S4C de a gogledd

Sampl cynrychiadol o ddata S4C (de):

```
IAWN
WI 'N FFAELU
OS FELLY LICIWN I WELD Y MUNUDAU O 'R CYFARFOD LLE Y PENDERFYNWYD HYNNY
PLIS
DYW E DDIM YN GALLU FFORDDIO SYMUD ALLAN
WI WEDI TSIECIO DDWYWAITH AC MAE PUM DEG TRI YN CYNNWYS NI
MAE 'R DDAU OHONOCH CHI 'DI SYRTHIO MEWN CARIAD
HAIA
FE SY 'N MUSKET - QUEER
WEL DW I WEDI CLYWED AMDANO FO
BETH TI 'N WNEUD AR OL GWAITH
HAIA BECS
WEL O O LEIAF
SO PETHAU RHYNGTO FI A BECA
DW I ANGEN FY BEAUTY SLEEP
DIAWL GWI RION
DIOLCH BYTH AM HYNNY
WI 'N NABOD RHYWUN SY 'N GWNEUD AMATEUR DRAMATICS
WASTRAFF
FALLE DYLAI MR WATKINS GYMRYD DROSTO
SYLWEDDOLAIS I 'NA AR OL FY BREAKDOWN I
MAE 'N RHAID BOD PASTA 'DA TI YMA
IAWN I CHI YNDY DACH CHI YN CAEL MYND OFF I CHWARAE HEBDDO I
TI 'N GW'BOD
RHO LONYDD I SEREN
YN DEG AR WEDDILL Y DOSBARTH
WNEWCH CHI DDELIO 'DAG E
MAE FE WEDI BOD YN FRAINT I ARWAIN Y BOIS AT Y FUDDUGOLIAETH 'MA
CAPTAIN MA 'N DDIGON BLIN 'DA FI FEL MAE HI
MAE E WEDI GWNEUD I FI SYLWEDDOLI
MAE 'N SIARAD AM FARDDONIAETH AR Y RADIO WEITHIAU
MAGGOT
O MR ROWLANDS YDY RWAN YN
WNEITH E DDIM HELPU
DY FRAWD
PARTI HENO
MAE HYNNA 'N BETH DA NAGYW E
BY' RHAI FI GA'L PREMIERE DIOLCH GRET
BETH TI 'N WATSIO
FI 'N MYND AM SLASH
NEFOEDD BETH SY 'DA TI I MEWN FAN HYN
ALLA I 'CH HELPU CHI
DDRWG GEN I MOD I 'N HWYR O 'N I 'N CAEL TRAFFERTH EFO 'R YSGRIFENYDDES
NEWYDD
SA I 'N GWELD BETH YW 'R BROBLEM WI 'N ATHRO DA
WEL NAWR BOD FI 'N GWYBOD BOD Y SIOE DDIM YN PANTS
OND WI WEDI AGOR Y GWIN NAWR
O GRET IE DIOLCH
FALLE DYLEN NI FYND ALLAN
CYN I 'R NORTH POLE DODDI
BYDDAI 'N GLIR 'DA FE 'TE
```

Sampl cynrychiadol o ddata S4C (gogledd):

SBIWCH LLANAST A 'R BATHROOM HEFYD
NESH I JYST BEICHIO CRIO FATHA FFWL
PA UN GAWN NI DWAD
WEL PAM NA DDEI DI DDIM A 'R CWBL I TY NI TA
DW I 'N BRYSUR
SA R'WBATH BLAW RHAI CAWS RHYFADD 'MA
AMSER YMA FORY BYDD PAWB YN CAPEL YN DISGWYL AMDANA CHDI
IA FALLE
TI 'DI ADOPTIO GAVIN 'TA BETH
JYST DOS I MEWN I TY
ADVERT YN Y PAPUR
TI 'N STYRIO JAC MURPHY
ANGIE 'N G'NEUD JOBAN DDA
OEDDACH CHI 'N YMWYBODOL FELLY BOD Y PLANT YN YFED ALCOHOL YN Y DAFARN
TOO RIGHT
MAE PARCH
I RHOI LLONYDD IDDYN NHW DE
MAE 'NA FWYO BYBS A CAFFIS YN Y DRE NA SIOPAU
GYNNOT TI OIL AR DY FOCH
TI 'N UFFAR O BISHYN CERI
TY' 'MA
RHY HWYR RWAN YNDYDY
A RWAN BO' VAL YN MYND MA' PETHAU 'N MYND I FOD YN FAIN ARNON NI
TI 'N TRIO DEUD BO' FI 'N FUDR NE' 'BATH
DOS A SAMOSA I CO BACH 'DE
NA DW I DDIM WEDI BOD YN GWNEUD Y FATH BETH
O'N
GWYCH
DW I 'M ISHO TRWBL
WEL DDEUDAIS I DW I I FOD YN GWEITHIO
MA' HWNNA 'N MYND AR 'N NERVES I WEITHIA'
OS ALLITH JAC 'NEUD O
DW I 'N MANIJO
POENI BE' 'NEITH O EFO 'I AMSER DW I FYDD O MOR BORED 'DE
RHAG OFN BE'
S'DIM LOT O BWYNT
NID DYNA GLYWAIS I
PA GI CHDI
YR UN WEN 'TA HONNO SY' EFO 'R WIGWAM EFFECT 'DE
DYNA FO DDEUD DWAD
MAE 'NA WAITH TREINIO AR HWN DE
OH NGWAS I
SIM ISHO MYND DROS BEN LLESTRI CHWAITH NAGO'S
FEDRAN NI 'M AROS - GORMOD O BAGGAGE
Y YLI BE' NA'I
AROS DI 'N FAN 'MA IAWN
A DYNA JINI YN DIGWYDD SON BOD HI 'N CHWILIO AM LOJAR
DOS ADRA 'DE
RHAID IDDO FO DDERBYN CHDI A 'R BABI BYDD
D'EUD Y GIWR O 'N I 'N GOBEITHIO MYND EFO DAD