META NET

White Paper Series Cyfres Papurau Gwyn

THE WELSH LANGUAGE IN THE DIGITAL AGE

Y GYMRAEG YN YR OES DDIGIDOL

Jeremy Evas

Springer

White Paper Series Cyfres Papurau Gwyn

THE WELSH LANGUAGE IN THE DIGITAL AGE

Y GYMRAEG YN YR OES DDIGIDOL

Jeremy Evas
Ysgol y Gymraeg, Prifysgol Caerdydd
School of Welsh, Cardiff University

Georg Rehm, Hans Uszkoreit
(Golygyddion, editors)

RHAGARWEINIAD

Mae'r papur gwyn hwn yn rhan o gyfres sy'n hyrwyddo gwybodaeth am dechnoleg iaith a'i photensial. Mae'n targedu addysgwyr, newyddiadurwyr, llunwyr polisi, gwleidyddion, cynrychiolwyr a chymunedau ieithyddol ac eraill. Mae faint o dechnoleg iaith sydd ar gael yn Ewrop a'r defnydd ohoni yn amrywio rhwng ieithoedd. O ganlyniad, mae'r camau gweithredu y mae eu hangen i gefnogi ymchwil bellach ac i ddatblygu technolegau iaith hefyd yn wahanol ar gyfer pob iaith. Mae'r camau gweithredu angenrheidiol yn dibynnu ar sawl ffactor, megis cymhlethdod yr iaith a nifer ei siaradwyr. Yn y gyfres hon o bapurau gwyn, mae META-NET, rhwydwaith o ragoriaeth a ariennir gan y Comisiwn Ewropeaidd, wedi cynnal dadansoddiad o'r adnoddau a thechnolegau ieithyddol cyfredol sydd ar gael. Canolbwyntiodd y dadansoddiad ar 23 iaith swyddogol Ewrop yn ogystal ag ieithoedd cenedlaethol a rhanbarthol pwysig eraill Ewrop. Mae canlyniadau'r dadansoddiad hwn yn awgrymu bod llawer o fylchau ymchwil sylweddol ar gyfer pob iaith. Bydd dadansoddiad mwy manwl-arbenigol ac asesiad o'r sefyllfa bresennol yn helpu i fanteisio i'r eithaf ar ymchwil ac i leihau risgiau. Fis Tachwedd 2011, roedd META-NET yn cynnwys 54 o ganolfannau ymchwil o 33 o wledydd Ewropeaidd sydd yn gweithio gyda rhanddeiliaid o'r economi (cwmnïau meddalwedd, darparwyr a defnyddwyr technoleg) asiantaethau llywodraethol, sefydliadau ymchwil, sefydliadau anllywodraethol, cymunedau ieithyddol a Phrifysgolion Ewropeaidd. Ar y cyd â'r cymunedau hyn, mae META-NET wrthi'n creu gweledigaeth gyffredin ar gyfer technoleg ac agenda ymchwil strategol ar gyfer Ewrop amlieithog yn 2020

PREFACE

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others. The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies in this white paper series. The analysis focuses on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of future research.

As of November 2011, META-NET consists of 54 research centres in 33 European countries. META-NET is working with stakeholders from economy (software companies, technology providers and users), government agencies, research organisations, non-governmental organisations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

Mae awdur y ddogfen hon yn ddiolchgar i awduron y papur gwyn ar yr Almaeneg [1] am ganiatâd i ailddefnyddio peth deunydd annibynnol-ar-iaith o'u dogfen ac i Laura Amey, John Judge, Jonathan Morris, Michael Speed, a Colin Williams am eu cymorth gyda'r papur hwn.

Ariannwyd datblygu'r papur gwyn hwn gan Raglen Seithfed Fframwaith a Rhaglen Cefnogi Polisi TG y Comisiwn Ewropeaidd o dan y contractau T4ME (Cytundeb Grant 249119), CESAR (Cytundeb Grant 271022), META-NET4U (Cytundeb Grant 270893) a META-NORD (Cytundeb Grant 270899).

The author of this document is grateful to the authors of the white paper on German for permission to re-use selected language-independent materials from their document [1], and to Laura Amey, John Judge, Jonathan Morris, Michael Speed, and Colin Williams for their help with this paper.

The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under the contracts T4ME (Grant Agreement 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899).

TABL CYNNWYS CONTENTS

Y Gymraeg yn yr Oes Ddigidol

THE WELSH LANGUAGE IN THE DIGITAL AGE

1

CRYNODEB GWEITHREDOL

Iaith yw'r prif ddull o gyfathrebu rhwng pobl. Mae'n ein galluogi i fynegi syniadau a theimladau, yn ein helpu i ddysgu ac addysgu, mae'n hanfodol ar gyfer byw. Dyma'r prif ddull o drosglwyddo diwylliant, ac mae'n symbol o'n hunaniaeth. O ystyried y globaleiddio sydd ohoni, mae gennym lawer o ddulliau o gyfathrebu'n hawdd gyda phobl o bob cwr o'r byd. Er enghraifft, mae'r technolegau gwybodaeth a chyfathrebu newydd wedi galluogi datblygu rhwydweithiau cymdeithasol sydd wedi annog a gwella rhyngweithio rhwng pobl o bron pob gwlad a diwylliant. Hefyd, yn ystod y blynyddoedd diwethaf, yr ydym wedi gweld symudiadau mawr o bobl tramor rhwng ein gwledydd, h.y. twristiaeth neu fewnfudo—sy'n creu'r angen am gyfathrebu ymysg gwahanol ieithoedd. Yn aml bydd y broblem cyfathrebu traws-ieithog hon yn cael ei datrys drwy ddefnyddio lingua franca. Yng ngwledydd Ewrop ceir enghraifft eglur o amrywiaeth ieithyddol a diwylliannol er gwaethaf y ffaith i Ewrop ymffurfio fwyfwy yn endid gwleidyddol ac economaidd penodol yn ystod y 60 mlynedd diwethaf. O ganlyniad, mae'n anochel ein bod yn wynebu heriau ieithyddol mewn bywyd bob dydd yn ogystal ag ym meysydd busnes, gwleidyddiaeth a'r gwyddorau.

Mae'n anochel ein bod yn wynebu heriau ieithyddol mewn bywyd bob dydd yn ogystal ag ym meysydd busnes, gwleidyddiaeth a'r gwyddorau

Bydd sefydliadau'r Undeb Ewropeaidd yn gwario tua biliwn Ewro y flwyddyn ar gynnal eu polisi amlieithrwydd, h.y. cyfieithu testunau a chyfieithu ar y pryd. Ar yr un pryd, mae Saesneg yn troi'n lingua franca wrth i'r sefydliadau Ewropeaidd gyfathrebu â'u dinasyddion. Yn y DU, fel enghraifft o hyn, mae gennym sefyllfa debyg. Gan fod llawer o wasanaethau cyhoeddus bellach yn cael eu darparu naill ai'n uniongyrchol neu'n anuniongyrchol drwy ddulliau technolegol, mae darparu a chofnodi dewis iaith, a'r dechnoleg iaith angenrheidiol i roi'r dewis hwn ar waith bellach yn fater o bwys am sawl rheswm. Mae'r cyfiawnhad mwyaf perthnasol yn ymwneud â hyrwyddo: dinasyddiaeth weithgar, tegwch o ran mynediad at wasanaethau meddygol, hygyrchedd ar gyfer y rhai, er enghraifft, sydd â nam ar eu golwg, a chynrychiolaeth ddemocrataidd ei hun.

Gall technoleg bontio rhwng gwahanol ieithoedd

O'i chyfuno â dyfeisiau deallus a rhaglenni, bydd technoleg iaith y dyfodol yn gallu helpu dinasyddion i siarad yn rhwydd â'i gilydd a gwneud busnes â'i gilydd, hyd yn oed os nad ydynt yn siarad iaith gyffredin. Mae hyn, yng nghyd-destun deddfwriaeth ieithyddol sydd newydd ei phasio sy'n effeithio ar y gwasanaethau cyhoeddus yng Nghymru, yn hollbwysig. Bydd atebion technoleg iaith yn y pen draw yn gwasanaethu fel pont unigryw rhwng gwahanol ieithoedd. Fodd bynnag, mae'r technolegau iaith ac offer prosesu lleferydd sydd ar gael ar y farchnad ar hyn o bryd (yn amrywio o systemau ateb cwestiwn i ryngwynebau iaith naturiol, gan gynnwys systemau cyfieithu ac offer crynhoi, ymhlith llawer o rai eraill), yn dal i fod heb gyrraedd y nod uchelgeisiol hwn. Mor gynnar â diwedd y 1970au, sylweddolodd yr Undeb Ewropeaidd berthnasedd arwyddocaol technoleg iaith fel grym

i yrru undod Ewropeaidd, a dechrau ariannu'r prosiectau ymchwil cyntaf o fewn y maes datblygol hwn. Ar yr un pryd, sefydlwyd prosiectau cenedlaethol ac awtonomig a oedd yn cynhyrchu canlyniadau gwerthfawr ond ni chafwyd gweithgarwch Ewropeaidd cydlynol. Mae prif dechnolegau iaith heddiw yn dibynnu ar ddulliau ystadegol nad ydynt yn fanwl gywir, nad ydynt yn gwneud defnydd o ddulliau ieithyddol dyfnach, rheolau a gwybodaeth. Er enghraifft, bydd brawddegau'n cael eu cyfieithu'n awtomatig gan gymharu brawddeg newydd yn erbyn miloedd o frawddegau a gyfieithwyd yn flaenorol gan fodau dynol. Mae ansawdd yr allbwn i raddau helaeth yn dibynnu ar faint ac ansawdd y corpws sampl sydd ar gael. Fodd bynnag, mae hyd yn oed y dull ystadegol lled-anghywir yn llawer mwy cynhyrchiol na llafur cyfieithydd unigol nad yw'n ymelwa ar dechnoleg o'r fath, neu'n ymelwa ar rannu amser-go-iawn â gwaith cyfieithwyr eraill drwy Gof Cyfieithu. Er y gall cyfieithu brawddegau syml yn awtomatig mewn ieithoedd a chanddynt symiau digonol o ddeunydd testun ddarparu canlyniadau defnyddiol, mae dulliau ystadegol bas yn rhwym o fethu yn achos ieithoedd a chanddynt gorff llawer llai o ddeunydd sampl, neu yn achos brawddegau a chanddynt strwythurau cymhleth. Dadansoddi priodweddau strwythurol dyfnach ieithoedd yw'r unig ffordd ymlaen os dymunir adeiladu rhaglenni sy'n perfformio'n dda ar draws ystod eang o ieithoedd. Felly, yr ateb i broblem cyfathrebu rhwng ieithoedd yw adeiladu technolegau galluogi allweddol. Er mwyn cyrraedd y nod hwn a chadw amrywiaeth ddiwylliannol ac ieithyddol Ewrop, yn gyntaf, mae angen dadansoddiad systematig o nodweddion ieithyddol holl ieithoedd Ewrop, a chyflwr presennol technoleg iaith i'w cefnogi. Dyma bwrpas y papur gwyn hwn sy'n ymwneud â'r Gymraeg.

Yr ateb i broblem cyfathrebu rhwng ieithoedd yw adeiladu technolegau galluogi allweddol

Fel y mae'r gyfres hon o bapurau gwyn yn ei ddangos, mae gwahaniaeth dramatig rhwng aelod-wladwriaethau Ewrop o ran aeddfedrwydd y gwaith ymchwil a pharodrwydd yr ymateb i adnabod a gweithredu atebion iaith. Un o gynigion a chasgliadau sy'n deillio o'r dystiolaeth o'r hyn sydd ar gael yw bod y Gymraeg yn un o ieithoedd yr UE sydd angen ymchwil bellach cyn bod technoleg iaith yn barod i'w defnyddio mewn sefyllfaoedd bob dydd yn eang, cyn bod yr iaith wedi'i normaleiddio ym myd technoleg a chyn bod y dechnoleg honno'n normaleiddio'r Gymraeg i'w llawn botensial. Tra bo technoleg iaith yn dechnoleg galluogi ac nid yn nod ynddi ei hun ar gyfer y person-yn-y-stryd, mae'n hanfodol bod ieithoedd llai fel y Gymraeg yn derbyn sylw dyledus neu bydd eu siaradwyr yn cael eu difreinio ymhellach.

Mae'n hanfodol bod ieithoedd lleiafrifol yn cael sylw dyledus, neu bydd eu siaradwyr yn cael eu difreinio ymhellach

Mewn ardal lle na fydd yw siaradwyr iaith yn dod i adnabod ei gilydd mewn modd ethnig neu fel arall, mae gan dechnoleg rôl fawr i'w chwarae o ran cofnodi dewis iaith dinasyddion. Byddai'r sefyllfa galluogi ddelfrydol hon yn gweld y dechnoleg yn galluogi asiantau'r wladwriaeth a sectorau eraill i gynnig gwasanaethau yn Gymraeg *mewn modd rhagweithiol*, gan y bydd dewis iaith dinasyddion eisoes yn hysbys iddynt. Er bod y sefyllfa ddeddfwriaethol yng Nghymru a ddisgrifir isod yn datblygu ei disgwrs ar gyfer y wladwriaeth i fod yn ddarparwr gwasanaethau o'r fath, bydd technoleg yn galluogi tegwch o ran darpariaeth iaith i bob dinesydd (e.e. cyfeirio galwadau ffôn Cymraeg yn awtomatig i staff sy'n siarad Cymraeg, cyfateb staff Cymraeg yn y gwasanaethau cymdeithasol/meddygol â defnyddwyr gwasanaeth sy'n siarad Cymraeg ac yn y blaen).

Bydd technoleg yn galluogi tegwch o ran darpariaeth iaith i bob dinesydd

2

IEITHOEDD MEWN PERYGL: HER I DECHNOLEG IAITH

Rydym yn dystion i chwyldro digidol sy'n effeithio cyfathrebu a'r gymdeithas mewn modd dramatig. Weithiau mae'r datblygiadau diweddar mewn technoleg gwybodaeth ddigidol a chyfathrebu yn cael eu cymharu â'r wasg argraffu a grëwyd gan Gutenberg. Beth gall y gymhariaeth hon ei ddweud wrthym am y dyfodol cymdeithas gwybodaeth Ewrop a'n hiaith ni yn benodol? Wedi i Gutenberg ddyfeisio'r wasg, cafwyd datblygiadau gwirioneddol mewn cyfathrebu a chyfnewid gwybodaeth trwy ymdrechion megis cyfieithu'r Beibl i ieithoedd brodorol.

Mae modd cymharu'r chwyldro digidol â gwasg argraffu Gutenberg

Yn y canrifoedd dilynol, datblygwyd technegau diwylliannol i drin prosesu iaith a chyfnewid gwybodaeth yn well:

- fe wnaeth safoni orgraffyddol a gramadegol y prif ieithoedd alluogi lledaenu syniadau gwyddonol a deallusol newydd yn gyflym
- fe wnaeth datblygu ieithoedd swyddogol greu posibilrwydd i ddinasyddion gyfathrebu o fewn ffiniau penodol (yn aml rhai gwleidyddol)
- mae addysgu a chyfieithu ieithoedd wedi galluogi cyfnewidfeydd ar draws ieithoedd; mae creu canllawiau golygyddol a llyfryddiaethol wedi sicrhau ansawdd deunydd print
- mae creu gwahanol gyfryngau megis papurau newydd, radio, teledu, llyfrau, a fformatau eraill wedi bodloni anghenion cyfathrebu gwahanol.

Yn yr un modd, yn ystod yr ugain mlynedd diwethaf, mae technoleg gwybodaeth wedi helpu i awtomeiddio a hwyluso ymhellach prosesu iaith a chyfnewid gwybodaeth:

- mae meddalwedd cyhoeddi bwrdd gwaith wedi cymryd lle teipio a chysodi;
- mae Microsoft PowerPoint, ac yn fwy diweddar, Prezzi wedi disodli tryloywderau taflunydd;
- mae e-bost yn caniatáu i ddogfennau gael eu hanfon a'u derbyn yn gyflymach na defnyddio peiriant ffacs;
- mae Skype a rhaglenni eraill yn darparu galwadau ffôn rhad ar y rhyngrwyd a chynnal cyfarfodydd rhithwir;
- mae fformatau amgodio sain a fideo yn ei gwneud yn hawdd cyfnewid cynnwys aml-gyfryngol;
- mae peiriannau chwilio ar y we yn darparu mynediad yn seiliedig ar allweddeiriau;
- mae gwasanaethau ar-lein megis Google Translate yn cynhyrchu cyfieithiadau cyflym ond bras;
- mae platffyrm cyfryngau cymdeithasol megis Facebook a Twitter yn hwyluso rhannu cyfathrebu, cydweithio, a gwybodaeth.

Er bod yr offer a'r rhaglenni hyn yn ddefnyddiol, nid ydynt eto yn gallu cynnal cymdeithas Ewropeaidd am-

lieithog hollol gynaliadwy y gall gwybodaeth a nwyddau lifo'n rhydd ynddi.

2.1 MAE FFINIAU IEITHYDDOL YN LLESTEIRIO CYMDEITHAS GWYBODAETH EWROP

Ni ellir rhagweld union ffurf a siâp cymdeithas gwybodaeth y dyfodol. Ond mae tebygolrwydd cryf bod y chwyldro mewn technoleg cyfathrebu yn dod â phobl sy'n siarad ieithoedd gwahanol at ei gilydd mewn ffyrdd newydd. Mae hyn yn rhoi pwysau ar unigolion i ddysgu ieithoedd newydd ac yn arbennig ar ddatblygwyr i greu rhaglenni technoleg newydd i sicrhau cyd-ddealltwriaeth ymysg siaradwyr gwahanol ieithoedd a mynediad at wybodaeth y gellir ei rhannu. Mewn gofod economaidd a gwybodaeth fyd-eang, mae mwy o ieithoedd, siaradwyr a chynnwys yn rhyngweithio'n gyflymach â mathau newydd o gyfryngau. Megis dechrau'r peth yw poblogrwydd cyfredol cyfryngau cymdeithas 'Gwe2.0' (megis Wikipedia, Facebook, Twitter, YouTube, ayb). Cynnwys sydd frenin a bellach, mae'r defnyddiwr yn rheoli i raddau na welwyd mo'u tebyg o'r blaen. Mae'r rhyddhau hwn rhag rheolaeth yn debyg i'r chwyldro mewn cyfalafiaeth argraffu a gynorthwyodd i ffurfio gwladwriaeth genedl fonolithig uniaith dair canrif yn ôl. Fodd bynnag, y tro hwn, *unigolion* sy'n rheoli, ac mae gan eu hieithoedd lwybr ar gyfer mynegiant a wadwyd iddynt hyd yma. A allai Gwe2.0, o'i chysylltu â thechnoleg iaith berthnasol fod yn gyfalafiaeth argraffu newydd a ddefnyddid yn hanesyddol fel grym homogeneiddio gan wladwriaethau cenedl, ond y tro hwn, *o blaid ieithoedd lleiafrifol*?

Heddiw, gellir trosglwyddo gigabeit o destun o gwmpas y byd mewn ychydig eiliadau cyn inni sylweddoli ei fod mewn iaith nad ydym yn ei deall. Yn ôl adroddiad gan y Comisiwn Ewropeaidd [2], bydd 57% o ddefnyddwyr y rhyngrwyd yn Ewrop yn prynu nwyddau a gwasanaethau mewn ieithoedd nad ydynt yn frodorol iddynt (Saesneg yw'r iaith dramor fwyaf cyffredin wedyn Ffrangeg, Almaeneg a Sbaeneg). Bydd 55% o ddefnyddwyr yn darllen cynnwys mewn iaith dramor ond dim ond 35% yn defnyddio iaith arall i ysgrifennu e-bost neu i bostio sylwadau ar y we. Ychydig flynyddoedd yn ôl, efallai y byddai'r Saesneg wedi bod yn lingua franca ar y we—roedd mwyafrif helaeth cynnwys y we yn Saesneg, o bosibl oherwydd iddi gael ei datblygu yn y lle cyntaf mewn gwledydd Saesneg eu hiaith. O ran cyfnod cynnar datblygu gwe-cynnwys ar gyfer ieithoedd RML, un rhagdybiaeth yw y gall hyn fod wedi digwydd oherwydd y meddylfryd a'r agweddau diglosig a ddisgrifir uchod.

> Mewn gofod economaidd a gwybodaeth fyd-eang, mae mwy o ieithoedd, siaradwyr a chynnwys yn rhyngweithio'n gyflymach â mathau newydd o gyfryngau

Yn ffodus, mae'r sefyllfa bellach wedi newid yn sylweddol. Mae cyfran y cynnwys ar-lein mewn ieithoedd Ewropeaidd eraill (yn ogystal â rhai Asia a'r Dwyrain Canol) wedi ffrwydro. Yn rhyfedd iawn, nid yw'r rhaniad digidol hollbresennol hwn o ganlyniad i ffiniau ieithyddol wedi cael llawer o sylw cyhoeddus, ac eto, mae'n codi cwestiwn pwysig iawn: Sut *yn union* y gall technolegau alluogi ieithoedd i ffynnu yn y gymdeithas gwybodaeth wedi'i rhwydweithio pan fydd cynifer ohonynt mewn perygl? Er enghraifft, mae Ethnologue [3] yn nodi bod tua 7,015 o ieithoedd yn y byd. Mae'r rhan fwyaf o'r rhain o dan risg ac ni fyddant yn cael eu trosglwyddo i genedlaethau'r dyfodol (mae'r fath drosglwyddo iaith yn un o ddangosyddion allweddol bywiogrwydd iaith). Mae'r fath risgiau yn destun adran nesaf y papur gwyn hwn.

Nid yw'r rhaniad digidol hollbresennol hwn o ganlyniad i ffiniau ieithyddol wedi cael llawer o sylw cyhoeddus

2.2 EIN HIEITHOEDD MEWN PERYGL

Er bod y wasg argraffu a'r gyfalafiaeth argraffu gysylltiedig wedi cynorthwyo i gyflymu cyfnewid gwybodaeth yn Ewrop [4], mae hefyd yn cyfrannu'n sylweddol at y broses lle mae llawer o ieithoedd Ewropeaidd yn colli nifer sylweddol o siaradwyr. Anaml iawn y byddai llawer o ieithoedd rhanbarthol neu leiafrifol yn cael eu hargraffu a chawsant eu cyfyngu i ffurfiau llafar, sydd yn ei dro yn cyfyngu ar sgôp eu defnyddio. Fe wnaeth y Gymraeg elwa ar iaith ysgrifenedig safonol a grëwyd gan yr Esgob William Morgan yn ei gyfieithiad o'r Beibl yn 1588. Ond sut y bydd y Gymraeg yn goroesi effaith y rhyngrwyd? Mae oddeutu 80 o ieithoedd yn Ewrop—dyma rai o'i hasedau diwylliannol cyfoethocaf a phwysicaf, [5] ac maent yn rhan hanfodol o'i model cymdeithasol unigryw. Er bod ieithoedd megis Saesneg a Sbaeneg yn debygol o oroesi yn y farchnad ddigidol sy'n dod i'r amlwg, gallai llawer o ieithoedd Ewropeaidd droi'n amherthnasol mewn cymdeithas sydd wedi'i rhwydweithio oni bai digon o gamau strategol yn cael eu cymryd. Byddai hyn yn gwanhau statws byd-eang Ewrop, ac yn mynd yn groes i'r nod strategol o sicrhau cyfranogiad cyfartal i bob dinesydd Ewropeaidd waeth beth yw eu hiaith. Mae amrywiaeth ieithyddol eang Ewrop gyda'i hasedau diwylliannol cyfoethocaf a phwysicaf. Yn ôl adroddiad UNESCO [6] ar amlieithrwydd, mae iaith yn gyfrwng hanfodol ar gyfer mwynhau hawliau sylfaenol, megis mynegiant gwleidyddol, addysg a chyfranogiad mewn cymdeithas.

O'r fan hon, symudir oddi wrth elfennau gwleidyddol, athronyddol, hanesyddol a strategol, i rai gweithredol. I ryw raddau, ac i orsymleiddio'n ddybryd gan esgeuluso'r cyd-destun cymdeithasol a amlinellwyd uchod, *gellid* portreadu iaith (wedi ei diriaethu ar ffurf gwrthrych y gellir ei fanipwleiddio) y tu allan i'r cyd-destun cymdeithasol hwnnw ac at ddibenion TG, fel 'dim byd mwy na chynnwys.' I grynhoi, mae gan dechnoleg a thechnoleg iaith ran fawr i'w chwarae wrth helpu bywydau pobl ddwyieithog ac i gael ei defnyddio fel offeryn cynllunio statws i newid agweddau dwfn tuag at y defnydd iaith 'L' mewn meysydd sydd hyd yn hyn wedi bod yn anghyfarwydd iddynt. Mae technoleg yn hollbresennol. Rhaid i amlieithrwydd fod yn hollbresennol mewn technoleg.

Sut y bydd y Gymraeg yn goroesi effaith y Rhyngrwyd?

2.3 MAE TECHNOLEG IAITH YN DECHNOLEG ALLWEDDOL AR GYFER GALLUOGI

Yn y gorffennol, bu ymdrechion mewn cadwraeth ieithyddol yn canolbwyntio ar ddysgu'r iaith a chyfieithu. Yn ôl un amcangyfrif, gwerth y farchnad Ewropeaidd ar gyfer cyfieithu, dehongli, lleoleiddio meddalwedd a globaleiddio gwefannau oedd €8.4 biliwn yn 2008 a disgwylir iddi dyfu o 10% y flwyddyn [7]. Eto mae'r ffigur hwn ond yn cwmpasu cyfran fechan o anghenion y presennol a'r dyfodol o ran cyfathrebu rhwng ieithoedd. Yr ateb mwyaf cymhellol a fydd yn sicrhau bod ieithoedd Ewrop yn cael eu defnyddio'n eang ac mewn sawl maes yw defnyddio technoleg briodol, yn union fel yr ydym yn defnyddio technoleg i ddatrys ein hanghenion cludiant, ynni ac anabledd ymhlith eraill. Gall technoleg iaith sy'n targedu pob math o gyfryngau ysgrifenedig a llafar helpu pobl i gydweithio, i gynnal busnes,

i rannu gwybodaeth ac i gymryd rhan mewn dadleuon cymdeithasol a gwleidyddol gan ddiystyru rhwystrau iaith a sgiliau cyfrifiadurol.

Mae technoleg yn hollbresennol. Rhaid i amlieithrwydd fod yn hollbresennol mewn technoleg

Yn aml bydd eisoes ar waith heddiw mewn modd anweledig y tu mewn i systemau meddalwedd cymhleth i'n helpu i: ddod o hyd i wybodaeth gyda pheiriant chwilio; gwirio sillafu a gramadeg mewn prosesydd geiriau; i weld argymhellion cynnyrch mewn siop ar-lein, i ddilyn cyfarwyddiadau llafar system llywio; i gyfieithu ar y we tudalennau trwy wasanaeth ar-lein. Mae technoleg iaith yn cynnwys nifer o raglenni craidd sy'n galluogi prosesau o fewn fframwaith mwy.

Pwrpas papurau gwyn iaith META-NET yw canolbwyntio ar ba mor barod y mae'r technolegau galluogi craidd ar gyfer pob iaith Ewropeaidd. Mae ar Ewrop angen technoleg iaith gadarn a fforddiadwy ar gyfer ei holl ieithoedd.

Mae ar Ewrop angen technoleg iaith gadarn a fforddiadwy ar gyfer ei holl ieithoedd

Er mwyn cynnal ein safle rheng flaen o ran arloesi bydeang, bydd angen ar Ewrop dechnoleg iaith, wedi'i theilwra i bob iaith Ewropeaidd, sy'n gadarn ac yn fforddiadwy a gellir ei hintegreiddio'n dynn mewn amgylcheddau meddalwedd allweddol. Heb dechnoleg iaith, ni fyddwn yn gallu sicrhau profiad rhyngweithiol, amlgyfryngol ac amlieithog i'r defnyddiwr yn effeithiol iawn yn y dyfodol agos.

2.4 CYFLEOEDD AR GYFER TECHNOLEG IAITH

Ym myd print, y datblygiad allweddol oedd galluogi dyblygu delwedd testun yn gyflym, a hynny drwy ddefnyddio gwasg argraffu ag iddi bŵer addas. Roedd yn rhaid i fodau dynol wneud y gwaith caled: chwilio, asesu, cyfieithu, a chrynhoi gwybodaeth. Bellach, gall technoleg iaith symleiddio ac awtomeiddio prosesau llif gwaith cyfieithu, creu cynnwys, a rheoli gwybodaeth. Gall hefyd rymuso rhyngwynebau llafar deallus ar gyfer offer electroneg yn y cartref, mewn peiriannau, cerbydau, cyfrifiaduron a robotiaid. Mae'n ddyddiau cynnar eto o ran datblygu rhaglenni masnachol a diwydiannol yn y byd go iawn, ond eto mae'r hyn a gyflawnwyd o ran ymchwil a datblygu yn creu cyfleoedd gwirioneddol. Er enghraifft, mae cyfieithu awtomatig (a ddisgrifir isod yn achos y Gymraeg) eisoes yn weddol gywir mewn meysydd penodol, ac mae rhaglenni arbrofol yn darparu gwybodaeth amlieithog ac yn rheoli gwybodaeth, yn ogystal â chreu cynnwys, mewn llawer o ieithoedd Ewropeaidd. Fel yn achos y rhan fwyaf o dechnolegau, datblygwyd y rhaglenni iaith cyntaf megis rhyngwynebau defnyddwyr seiliedig-ar-lais a systemau deialog ar gyfer meysydd arbenigol (er enghraifft, meysydd meddygol cul, ar gyfer cymryd nodiadau), ac yn aml cyfyngedig yw eu perfformiad. Fodd bynnag, mae marchnadoedd enfawr yn y diwydiannau addysg ac adloniant ar gyfer integreiddio technolegau iaith i mewn i gemau, pecynnau addysg-adloniant, llyfrgelloedd, amgylcheddau efelychu a rhaglenni hyfforddi. Gwasanaethau gwybodaeth symudol, meddalwedd cyfrifiadurol ben-bwrdd, e-ddysgu amgylcheddau dysgu cyfunol, offer hunanasesu a meddalwedd canfod llên-ladrata (a ddefnyddir i ganfod llên-ladrata gwaith wedi cyflwyno aseiniadau myfyrwyr) dyma rai yn unig o'r meysydd y gall technoleg iaith chwarae rôl bwysig ynddynt. Mae poblogrwydd rhaglenni cyfryngau cymdeithasol megis Twitter a Facebook yn awgrymu angen

am dechnolegau iaith soffistigedig sy'n gallu monitro postiadau, crynhoi trafodaethau, awgrymu tueddiadau barn, canfod ymatebion emosiynol, nodi tor-hawlfraint neu dracio camddefnydd. Mae'r un peth yn wir ar gyfer prosiectau ymchwil 'data mawr' [8] a rhai sy'n olrhain tueddiadau, fel y rhai a ddefnyddir gan y gwasanaethau diogelwch i ganfod y posibilrwydd o aflonyddwch cymdeithasol a geiriau 'casineb'. Mae technoleg Iaith yn helpu i oresgyn 'anabledd' (fel y byddai rhai yn ei alw) amrywiaeth ieithyddol. Mae technoleg iaith yn gyfle gwych ar gyfer yr Undeb Ewropeaidd. Gall helpu i fynd i'r afael â mater cymhleth amlieithrwydd yn Ewrop—sef y ffaith bod gwahanol ieithoedd yn cydfodoli'n naturiol mewn busnesau, sefydliadau ac ysgolion. Fodd bynnag, mae angen i ddinasyddion gyfathrebu ar draws ffiniau iaith Marchnad Gyffredin Ewrop, a gall technoleg iaith helpu i oresgyn y rhwystr olaf hwn, wrth gefnogi defnyddio ieithoedd unigol mewn modd rhydd ac agored. Gan fwrw golwg ymhellach i'r dyfodol, bydd technoleg iaith amlieithog arloesol yn Ewrop yn darparu meincnod ar gyfer ein partneriaid byd-eang pan fyddant yn dechrau cefnogi eu cymunedau amlieithog eu hunain.

> Mae technoleg Iaith yn helpu i oresgyn 'anabledd' (fel y byddai rhai yn ei alw) amrywiaeth ieithyddol

Er bod technoleg iaith wedi gwneud cynnydd sylweddol yn ystod y blynyddoedd diwethaf, mae cyflymder y cynnydd technolegol a'r arloesi sydd ohoni yn rhy araf o ran cynhyrchion. Fel arfer, bydd technolegau a ddefnyddir yn eang, fel gwirwyr sillafu a gramadeg mewn prosesyddion geiriau yn uniaith, ac maent ond ar gael ar gyfer llond llaw o ieithoedd. Mae gwasanaethau cyfieithu awtomatig ar-lein, er yn ddefnyddiol ar gyfer cynhyrchu brasamcan rhesymol o gynnwys dogfen yn gyflym, yn llawn anawsterau pan fydd angen cyfieithiadau hynod gywir ac yn gyflawn. Mae cyflymder y cynnydd sydd ohoni ym myd technoleg iaith yn rhy araf. Oherwydd cymhlethdodau iaith ddynol, mae darparu ar gyfer modelu cyfrifiadurol ein hieithoedd a'u profi yn y byd go iawn yn ofyn mawr o ran amser a chost ac mae gofyn am ymrwymiadau ariannu parhaus. Felly, rhaid i Ewrop gynnal ei rôl arloesol wrth wynebu'r heriau technolegol a grëir gan gymuned amlieithog trwy ddyfeisio dulliau newydd i gyflymu datblygiad ar draws y map.

2.5 CAFFAEL IAITH MEWN BODAU DYNOL A PHEIRIANNAU

Er mwyn dangos sut mae cyfrifiaduron yn trin iaith a pham y mae'n anodd eu rhaglennu i brosesu gwahanol ieithoedd, edrychwn yn gryno ar y ffordd y mae pobl yn caffael iaith gyntaf ac ail iaith, ac yna gweld sut y mae systemau technoleg iaith yn gweithio. Mae babanod yn caffael iaith trwy ryngweithio ieithyddol a thrwy wrando ar y rhyngweithio go iawn rhwng eu rhieni, brodyr a chwiorydd ac aelodau eraill o'r teulu. O'r adeg y byddant yn dathlu eu pen-blwydd yn ddwy oed, bydd plant yn cynhyrchu eu geiriau cyntaf ac ymadroddion byr. Mae hyn yn bosib dim ond oherwydd bod gan fodau dynol ragdueddiad genetig i ddynwared ac yna i resymoli'r hyn y maent yn ei glywed. Mae dysgu ail iaith wrth droi'n hŷn yn gofyn am ymdrech fwy gwybyddol, yn bennaf oherwydd nad yw'r person wedi'i 'foddi' mewn cymuned iaith o siaradwyr brodorol. Fel arfer, yn yr ysgol, bydd ieithoedd tramor yn cael eu caffael drwy ddysgu strwythur gramadegol, geirfa a sillafu drwy ddefnyddio driliau sy'n disgrifio gwybodaeth ieithyddol o safbwynt rheolau haniaethol, tablau ac enghreifftiau. Mae pobl yn caffael sgiliau iaith mewn dwy ffordd wahanol: dysgu o enghreifftiau a dysgu'r rheolau iaith sylfaenol. Gan symud yn awr at dechnoleg iaith,

mae'r ddau brif fath o systemau yn 'caffael' galluoedd iaith mewn modd tebyg. Mae dulliau ystadegol (neu rai a 'yrrir' gan ddata) yn cael gwybodaeth ieithyddol o gasgliadau helaeth o destunau enghreifftiol diriaethol neu 'gorpora'. Er ei bod yn ddigonol i ddefnyddio testun uniaith at ddibenion hyfforddi, dyweder, gwirydd sillafu, rhaid i destunau cyfochrog mewn dwy neu fwy o ieithoedd fod ar gael ar gyfer hyfforddi system cyfieithu awtomatig. Mae'r algorithm dysgu peirianyddol wedyn yn 'dysgu' patrymau o ran sut y bydd geiriau, ymadroddion byr a brawddegau cyflawn yn cael eu cyfieithu. Fel arfer, mae'r dull ystadegol hwn yn gofyn am filiynau o frawddegau, a hynny er mwyn rhoi hwb i ansawdd perfformiad. Dyma un rheswm pam y bydd darparwyr peiriannau chwilio yn awyddus i gasglu cymaint o ddeunydd ysgrifenedig ag y bo modd. Mae cywiro sillafu mewn prosesyddion geiriau, a gwasanaethau megis Google Search neu Google Translate, i gyd yn dibynnu ar ddulliau ystadegol. Mantais fawr y dull ystadegol yw bod y peiriant yn dysgu yn gyflym mewn cyfres barhaus o gylchoedd hyfforddi. Dull arall o fynd i'r afael â thechnoleg iaith a chyfieithu awtomatig yn benodol, yw adeiladu systemau yn seiliedig ar reolau. Yn gyntaf, rhaid i arbenigwyr ym meysydd ieithyddiaeth, ieithyddiaeth gyfrifiadurol a gwyddoniaeth gyfrifiadurol amgodio dadansoddiadau gramadegol (rheolau gramadegol) a llunio rhestrau geirfaol (geiriaduron). Mae hyn yn gofyn am lawer iawn o amser a llafur dwys. Mae rhai o'r systemau cyfieithu awtomatig ar sail rheolau wedi eu datblygu'n gyson am fwy na 20 mlynedd. Mantais fawr systemau sy'n seiliedig ar reolau yw bod gan yr arbenigwyr reolaeth fanylach dros sut y bydd yr iaith yn cael ei phrosesu. Mae hyn yn ei gwneud yn bosibl i gywiro camgymeriadau yn y meddalwedd mewn modd systematig ac i roi adborth manwl i'r defnyddiwr, yn enwedig pan fydd systemau seiliedig ar reolau yn cael eu defnyddio ar gyfer dysgu iaith. Fodd bynnag, oherwydd cost uchel y gwaith hwn, datblygwyd technoleg iaith seiliedig ar reolau ond ar gyfer ychydig ieithoedd mawr hyd yn hyn. Gan fod cryfderau a gwendidau systemau ystadegol a rhai seiliedig ar reolau yn dueddol o ategu ei gilydd, mae'r ymchwil gyfredol yn canolbwyntio ar ddulliau hybrid sy'n cyfuno'r ddwy fethodoleg. Mae'r posibiliadau o beiriant peiriant cyfieithu *tridarn* ar gyfer y Gymraeg (sy'n seiliedig ar reolau, yn seiliedig ar ystadegau ac yn seiliedig ar enghreifftiau), ynghyd â chof cyfieithu rhannu cyfieithiadau yn eang mewn amser go iawn yn cael eu trafod isod. Dylid nodi, fodd bynnag, bod y dulliau hybrid hyn hyd yn hyn wedi bod yn llai llwyddiannus mewn rhaglenni diwydiannol nag yn y labordy ymchwil. Mae'r ddau brif fath o systemau technoleg iaith yn caffael iaith yn yr un modd ag y bydd bodau dynol yn ei wneud. Fel yr ydym wedi ei weld yn y bennod hon, mae llawer o raglenni a ddefnyddir yn eang yn y gymdeithas gwybodaeth heddiw yn dibynnu'n helaeth ar dechnoleg iaith, yn enwedig yng ngofod economaidd a gwybodaeth Ewrop. Er bod y dechnoleg hon wedi gwneud cynnydd sylweddol yn ystod y blynyddoedd diwethaf, mae potensial mawr o hyd i wella ansawdd systemau technoleg iaith. Yn y penodau nesaf, disgrifir rôl y Gymraeg yng nghymdeithas gwybodaeth Ewrop a'r byd, ac asesir ei chefndir cymdeithasol-ieithyddol a chyflwr cyfredol technoleg iaith ar ei chyfer.

3

Y GYMRAEG YN Y GYMDEITHAS GWYBODAETH EWROPEAIDD

3.1 FFEITHIAU CYFFREDINOL

Mae i'r Gymraeg, (iaith Geltaidd sy'n gysylltiedig â Llydaweg, Cernyweg, Gwyddeleg, Gaeleg yr Alban a Manaweg) hanes tebyg i lawer arall o ieithoedd rhanbarthol neu leiafrifol eraill ar draws Ewrop yn sgil yr oleuedigaeth, hy polisïau canoli'r wladwriaeth yn arwain at ddirywiad demograffig sydyn ar ddechrau'r ugeinfed ganrif, wedi'i ddilyn gan arafu yn y gostyngiad yn y 1970au. Yn achos y Gymraeg, roedd canlyniadau'r cyfrifiad dengmlyneddol diweddaraf (2011) [9] yn dangos bod 19% o boblogaeth Cymru (562,000) yn nodi eu gallu i siarad Cymraeg, gostyngiad mewn niferoedd absoliwt o 20,400 ar y cyfrifiad blaenorol (2001). Dylid nodi, fodd bynnag bod Cyfrifiad 2001 yntau wedi dangos cynydd o 80,000 person ar gyfrifiad 1991. Byddai felly'n ymddangos, o'r sampl graddfa fawr a ddarperir gan y Cyfrifiad, bod y Gymraeg mewn sefyllfa lawer mwy diogel o ran ei ffigurau demograffig lefel uchaf na llawer o'r ieithoedd eraill yn y byd. Ond wrth gwrs, gall ffigurau lefel uchel o'r fath ond fod yn arwynebol. Mae'r darlun ei hun yn fwy cymhleth o lawer [10]. Mae'r cymhlethdod hwn yn amlygu ei hun ym mhroffil oedran siaradwyr Cymraeg sydd wrthi'n esblygu. Er enghraifft, o blith y grŵp oedran 10-14, cofnodwyd bod 42.2% yn siarad Cymraeg yng nghyfrifiad 2011, sy'n uwch na'r ffigurau canrif ynghynt yn 1911 (ac yn llawer uwch na'r 16.2% o'r rhai hyn na 65 mlwydd oed a nodododd eu bod yn siarad Cymraeg yn 2011).Yn wir, mae'r dadansoddiad diweddaraf o ffigurau Cyfrifiad 2011 ar gyfer y Gymraeg [11] yn dangos bod 'the number of children speaking Welsh is more than twice that of those aged 16-64 and the over 65s'. hefyd cyhoeddodd Bwrdd yr Iaith Gymraeg 'amcangyfrif rhesymol' [12] y gallai fod 110,000 o siaradwyr Cymraeg yn byw yn Lloegr. Mae goblygiadau cadarnhaol technoleg iaith a thechnoleg yn gyffredinol ar gyfer defnydd iaith a chynnal a chadw cymunedau ieithyddol gwasgaredig o'r fath yn sylweddol. Mae'r sianel teledu Cymraeg, S4/C, yn darlledu yn Lloegr (a gweddill y DU yn Gymraeg). O'r *holl* siaradwyr Cymraeg hynny (100%) mae 44.9% yn byw mewn cartrefi lle mae pawb yn siarad Cymraeg (ffigurau Cyfrifiad 2001—nid yw'r dadansoddiad cyfatebol ar gyfer ffigurau 2011 ar gael adeg llunio'r papur gwyn hwn). Mae'r ystadegau hyn yn golygu bod bywydau teuluol llawer iawn o siaradwyr Cymraeg yn ddwyieithog; mae hyn yn achosi heriau penodol o ran technoleg iaith y gellir newid iaith ei rhyngwyneb a hynny yng nghyd-destun cynllunio ieithyddol statws (h.y. sut y gall pobl sydd â gwahanol alluoedd ieithyddol yn yr un teulu neu uned yn y gweithle rannu cyfrifiadur os yw ei ryngwyneb yn yr iaith 'L' (statws diglosig is). Felly, mae'n hanfodol bwysig ar gyfer cynllunio ieithyddol cymwysedig bod cymaint â phosibl o'r dechnoleg y bydd pobl yn ei defnyddio yn yr ysgol a'r tu hwnt ar gael yn rhad ac am ddim yn Gymraeg, a'i bod yn cael ei gweithredu yn y sefydliadau a'r rhwydweithiau sy'n cael eu defnyddio gan segmentau arbennig penodol. Mae darparu dadansoddiad ystadegol o allu a defnydd y Gymraeg yn ôl oedran y tu hwnt i gwmpas y papur gwyn hwn. Fodd

bynnag, tâl inni nodi po fwyaf rhugl y bydd siaradwr yn y Gymraeg, po fwyaf tebygol ydyw y bydd y siaradwr hwnnw i ddefnyddio'r iaith bob dydd.

Mae goblygiadau cadarnhaol technoleg iaith a thechnoleg yn gyffredinol ar gyfer defnydd iaith a chynnal a chadw cymunedau ieithyddol gwasgaredig yn sylweddol

Nododd Arolygon Defnydd Iaith Gymraeg Bwrdd yr Iaith Gymraeg gynt (2004-2006) (dyma'r ffigurau defnydd iaith fwyaf diweddar sydd ar gael ar y lefel genedlaethol), o'r 588,000 o bobl yr amcangyfrifwyd eu bod yn siarad y Gymraeg, 58% (317,000) a oedd yn ystyried eu hunain yn rhugl yn yr iaith. Dywedodd 76% o'r siaradwyr rhugl eu bod yn siarad Cymraeg bob dydd, a Chymraeg oedd iaith sgwrs ddiweddaraf 59% o'r siaradwyr rhugl. Yn ystod oes Bwrdd statudol yr Iaith Gymraeg (1993-2012), esblygodd y drafodaeth ym maes Cynllunio Ieithyddol o awydd gorsyml dim ond i gynyddu'r niferoedd sy'n gallu siarad Cymraeg i strategaeth newid ymddygiad fwy soffistigedig yn defnyddio gwersi a ddysgwyd oddi wrth waith Hybu Iechyd y Gwasanaeth Iechyd Gwladol (cf prosiect Twf i berswadio rhieni i siarad Cymraeg â'u plant lle nad oeddent yn meddu ar ddigon o hyder i wneud hynny) ac i gynyddu defnyddio'r Gymraeg mewn sefyllfaoedd diglosig 'H' drwy brosiect o'r enw *Mae gen ti ddewis....* Mae *defnydd* yn hytrach na chynyddu niferoedd hefyd yn britho dogfennau strategaeth mwy diweddar Llywodraeth Cymru.

Mae patrymau ymddygiad ieithyddol wedi'u gwreiddio'n ddwfn mewn seicoleg gymdeithasol, hunan-ganfyddiadau, hunanhyder a chanfyddiadau o hunan-effeithiolrwydd ieithyddol (waeth beth yw rhuglder gwirioneddol siaradwr Cymraeg)

Mae patrymau ymddygiad ieithyddol wedi'u gwreiddio'n ddwfn mewn seicoleg gymdeithasol, hunan-ganfyddiadau, hunanhyder a chanfyddiadau o hunan-effeithiolrwydd ieithyddol [13] (waeth beth yw rhuglder gwirioneddol siaradwr Cymraeg). Mae'r elfennau hyn yn codi dro ar ôl tro yn yr ymchwil a gomisiynwyd gan Fwrdd yr Iaith Gymraeg gynt. Yn gryno, mae canfyddiad unigolyn gyfystyr â'i realiti goddrychol ei hun ac, ar y cyd â nifer o ffactorau eraill, bydd bod dynol yn gweithredu o fewn y paramedrau y bydd ei hunangred un yn eu creu. Mae hyn yn arbennig o berthnasol wrth ystyried y defnydd o dechnoleg iaith.

3.2 NODWEDDION Y GYMRAEG

Mae nodweddion ieithyddol cynhenid sy'n gwneud y Gymraeg yn annhebyg i lawer o ieithoedd eraill y gyfres hon papur gwyn (heblaw am ei chyfnither, y Wyddeleg). Gall hyn wneud datblygu, ac yn benodol, croesffrwythloni technoleg iaith yn fwy o her na, dyweder, rhwng Ffrangeg, Sbaeneg a Saesneg.

Mae 29 o lythrennau yn y wyddor Gymraeg (defnyddir sgript Rufeinig). Nid yw'r Gymraeg yn defnyddio 'x' neu 'z', a a defnyddir 'j' fel arfer mewn geiriau benthyg o'r Saesneg). Mae'r iaith yn ymelwa ar gefnogaeth Unicode lawn, ac felly, fel rheol gyffredinol, ni ddylai fod unrhyw broblem o ran dangos y nodau a ddefnyddir mewn unrhyw raglen sy'n gydnaws ag Unicode.

Un o brif hynodion y Gymraeg yw ei bod yn defnyddio deugraffau (dau symbol i gynhyrchu sain benodol). Mae'r rhain fel a ganlyn ch, dd, ff, ng, ll, mh, nh, ph, rh, th. Mae hyn yn cynnig her i'r technolegwyr hynny sy'n gweithredu, er enghraifft, didoli mewn cronfeydd data, gan fod 'Llandeilo' yn dod ar ôl 'Luton' (ill dau yn enwau lleoedd).

Mae'r Gymraeg, fel ei chyfnither y Wyddeleg, yn iaith ffurfdroadol sy'n golygu bod ei ffurfiau ieithyddol yn

newid gan ddibynnu ar (er enghraifft) amser, nifer, a pherson. Cymerwch, er enghraifft y ferf reolaidd 'canu'. Ar ei ffurf lenyddol, gryno, gallai'r ferf gael ei rhedeg fel a ganlyn:

- Canaf
- Ceni
- Cân
- Canwn
- Canwch
- Canant

Fodd bynnag, mae'r bwlch rhwng Cymraeg ysgrifenedig ffurfiol a llafar yn eang. Mae'r iaith lafar yn tueddu tuag at ddefnyddio strwythur mwy cwmpasog (nad yw ei hun heb ei broblemau o ganlyniad i amrywiadau tafodieithol sylweddol yn y strwythur cwmpasog hwnnw). Er enghraifft, yn y ffurf beriffrastig, byddid yn cyfleu 'I am singing/I sing' fel a ganlyn (gan ddibynnu ar y ffurf y bydd y siaradwr yn ei mabwysiadu):

- Dwi'n canu
- Rwy'n canu
- Rwyf yn canu
- Rydw i'n canu
- Fi'n canu (ffurf wedi'i stigmateiddio, ond yn dod yn fwy cyffredin)

Mae'r un amrywiad yn bodoli hefyd ar gyfer personau eraill. Mae diffyg iaith lafar safonol yn achosi problemau, er enghraifft, ar gyfer datblygu systemau adnabod lleferydd, neu yn wir cyfieithu awtomatig neu chwilio ffonig. Rhaid hefyd i'r gronfa ddata lemateiddio berthnasol ddelio â'r broblem hon.

Nodwedd arall o'r ieithoedd Celtaidd yw eu system o dreigladau i gytseiniaid ar ddechrau geiriau. Bydd naw llythyren yn treiglo yn y Gymraeg, a cheir tri math o dreiglad, fel y dangosir yn y tabl isod. Fodd bynnag, mae angen i dechnoleg iaith ystyried y rheolau sy'n achosi i'r treigladau hyn ddigwydd. Er enghraifft, gall y gair 'a' olygu 'yn ogystal â (and)', neu gall fod yn eiryn cyn-ferfol. Yn y lle cyntaf, byddai'n achosi treiglad llaes, yn yr ail, dreiglad meddal. Bydd gwrthrych ffurf gryno'r ferf hefyd yn cymryd treiglad meddal yn y Gymraeg. Felly, er enghraifft, mae i 'Gwelodd fachgen' ystyr wahanol i 'Gwelodd bachgen', a'r treiglad sy'n cyfleu hyn. Bydd enwau lleoedd yn cymryd treiglad trwynol ar ôl 'yn', ond gall 'yn' hefyd fod yn eiryn cyn-ferfol. Sut y bydd technoleg iaith yn gallu gwahaniaethu rhwng y sefyllfaoedd hyn? Mae hyn, wrth gwrs, hefyd yn berthnasol ar gyfer y cyfieithu awtomatig a drafodir yn ddiweddarach yn y papur gwyn hwn.

Er bod technoleg iaith wedi gwneud cynnydd sylweddol yn ystod y blynyddoedd diwethaf, mae cyflymder cyfredol cynydd ac arloesi cynnyrch technolegol yn araf ar gyfer llawer o ieithoedd llai Ewrop. Felly, rhaid i Ewrop gynnal ei rôl arloesol wrth wynebu'r heriau technolegol a grëir gan gymuned amlieithog trwy ddyfeisio dulliau newydd i gyflymu datblygiad ar draws y map. Gallai'r rhain gynnwys datblygiadau a thechnegau megis torfoli cyfrifiadurol. Yn achos penodol y Gymraeg, mae rhai heriau y mae angen eu goresgyn wrth ddatblygu technoleg iaith fel a ganlyn, e.e. mae amrywiaeth tafodieithol yn y Gymraeg yn fawr, er bod siaradwyr pob tafodiaith yn deall ei gilydd. Hyd nes dyfodiad S4/C, y sianel deledu Gymraeg yn 1982, sydd wedi hwyluso dealltwriaeth fewnol yn y gymuned lleferydd Gymraeg, clywyd llawer o gwynion anecdotaidd ynglŷn â'r diffyg gallu i ddeall tafodieithoedd Cymru. Mae hyn yn cynnig her amlwg i dechnoleg, megis adnabod llais. Mae Cymru wedi elwa o iaith ysgrifenedig safonol ers cyfieithiad o'r Beibl gan yr Esgob William Morgan yn 1588. Gwnaeth hyn lawer i warchod undod iaith, lle, er enghraifft, mae ieithoedd eraill, megis y Llydaweg, wedi ffragmenteiddio'n dafodieithoedd gwahanol. Mae crefydd wedi chwarae, hyd yn ddiweddar iawn, rôl gymdeithasegol fawr mewn cadwr-

	Meddal	Trwynol	Llaes
T	D	NH	TH
C	G	NGH	CH
P	B	MH	PH
B	F	M	-
D	DD	N	-
G	[Diflannu]	NG	-
M	F	-	-
RH	R	-	-
LL	L	-	-

aeth ieithyddol yng Nghymru, pan nad oedd y wladwriaeth mor barod i annog dwyieithrwydd ag y mae yn awr.

Gellid rhannu rhai cydrannau a'u hailddatblygu rhwng ieithoedd

Fodd bynnag, mae'r iaith ysgrifenedig safonol a ddisgrifir uchod yn wahanol iawn i Gymraeg llafar (yn ei holl ffurfiau tafodieithol). Felly, er enghraifft, yn achos defnyddio offer trosi llais yn destun, pa offer technoleg iaith a fyddai'n ofynnol er mwyn trawsgrifio'r ymadroddion i greu cofnod ffurfiol mewn Cymraeg ysgrifenedig o'r cyfarfod hwnnw (heb waith ôl-golygu sylweddol)? I ddangos y pwynt hwn, nododd yr Athro Bobi Jones [14] sydd wedi ysgrifennu'n helaeth ar faterion y Gymraeg, yng nghyd-destun Cymraeg ffurfiol 'Yr ydym yn siarad iaith ein mamau, ac yn ysgrifennu iaith neiniau neiniau ein hen neiniau!' Ar hyn o bryd, prin yw'r ddarpariaeth ffurfiol ar gyfer hyfforddi siaradwyr Cymraeg fel ieithyddion cyfrifiadurol. Mae'r setiau angenrheidiol o sgiliau technolegol, rhaglennu Cymraeg cywir ac ieithyddiaeth gyfrifiadurol yn brin. Yn aml bydd yn anodd penodi staff i swyddi *creu* cynnwys hyd yn oed (er enghraifft, ar gyfer gwefannau) a chanddynt ddigon o sgiliau iaith ac *ysgrifennu copi*.

Ceir arbedion mewn costau, cynnydd mawr o ran cysondeb a gostyngiad o ran cyfieithu ailadroddus drwy rannu Cofion Cyfieithu

Mae prosiectau Ewrop-eang megis META-NET yn dangos yr arbedion maint mawr y gellir eu canfod drwy fynd i'r afael â phroblemau tebyg llawer o ieithoedd mewn un lle. Yn ddiau, dyma gyfle i'r Gymraeg, ac ieithoedd llai eraill. Ac er nad yw llawer o ieithoedd Ewrop yn ramadegol debyg, gellid rhannu rhai cydrannau a'u hailddatblygu rhwng yr ieithoedd hynny *sy'n* gysylltiedig. Dylid defnyddio hefyd lawer o dechnolegau generig ac, o bosibl, eu rhannu eu pensaernïaeth, ar gyfer ieithoedd Ewrop, megis:

- Cofion cyfieithu sy'n seiliedig ar gwmwl
- Cyfieithu awtomatig hybrid (wedi ei gysylltu â'r cofion cyfieithu a ddisgrifir yn fanylach yn y papur gwyn hwn)

O ran y cyfleoedd ar gyfer y Gymraeg y gall technoleg iaith eu darparu, gellir ystyried:

- Y cyfalaf cymdeithasol a allai ddeillio o brosiectau torfoli yn y sector gwirfoddol, a'r cyfalaf cymdeithasol a ddaw o greu cynnwys ar y cyd yn y Gymraeg
- Arbedion mewn costau, cynnydd mawr o ran cysondeb a gostyngiad o ran cyfieithu ailadroddus drwy

rannu Cofion Cyfieithu yn y sector cyhoeddus ar raddfa fawr

- Gall synthesis lleferydd ddarparu cymorth ar gyfer y rhai sy'n dysgu Cymraeg nad oes ganddynt fynediad bob dydd i siaradwyr yr iaith
- O ran yr agenda cynhwysiant, bydd darllenwyr sgrin iaith Gymraeg yn cynyddu mynediad i'r rhai sydd â nam ar eu golwg i gynnwys a gwasanaethau iaith Gymraeg.

3.3 DATBLYGIADAU DIWEDDAR

Bu i athroniaeth cenedlaetholdeb rhamantaidd yn sgil yr oleuedigaeth wedi amlygu ei hun wrth greu gwladwriaethau cenedl monolithig sydd, er mwyn bod yn fonolithig, wedi gweithredu polisïau canoli, unffurfio a oedd yn gostwng (neu mewn rhai achosion, yn dileu) amrywiaeth ieithyddol mewnol. Mae agweddau sosioeicolegol y diglosia a grëwyd o'r herwydd ar gyfer technoleg a swyddogaethau 'modern' hefyd wedi eu hesbonio. Fodd bynnag, yn ystod traean olaf yr 20 fed datblygodd mudiad byd-eang o blaid *glocaleiddio*, a oedd yn rhoi pris mawr ar elfennau lleol ond eto yn meddu ar fyd-olwg byd eang. Gwelodd yr adfywiad hwn gynnydd mewn diddordeb mewn ieithoedd rhanbarthol neu leiafrifol. Yn achos y Gymraeg, fe wnaeth anufudd-dod sifil eang ledaenu a phrotest ddi-drais (yn bennaf, yn y blynyddoedd cynharach, ar ran Cymdeithas y Gymraeg [15]) sicrhau consesiynau gan y Llywodraeth. Cafwyd deddfwriaeth arall hefyd yn sgil hynny:

- Deddf yr Iaith Gymraeg (1967), a ganiataodd i weinidogion ragnodi fersiynau swyddogol o ffurflenni yn y Gymraeg, defnydd cyfyngedig o'r Gymraeg yn y system llysoedd a nifer o ddarpariaethau eraill.
- Deddfau Darlledu (1981 a 1982) a arweiniodd at sefydlu S4/C, y sianel deledu Gymraeg. Mae S4/C bellach yn ddarlledwr aml-lwyfan digidol arloesol sy'n cynnwys cynulleidfaoedd yn ei rhaglenni trwy gyfryngau cymdeithasol.
- Deddf Addysg 1988 a wnaeth y Gymraeg yn bwnc craidd yn y Cwricwlwm Cenedlaethol yng Nghymru (nid oedd yr iaith hyd hynny wedi'i dysgu'n orfodol yn y system ysgolion, ac nid oedd llawer o ysgolion yn ei dysgu o gwbl).
- Bwrdd Ymgynghorol yr Iaith Gymraeg (1988) a sefydlwyd i wneud argymhellion i Weinidogion ynghylch priodoldeb deddfu ar gyfer y Gymraeg. Fe achosodd argymhelliad o'r fath ddrafftio Deddf yr Iaith Gymraeg 1993. Mae'r Ddeddf honno, sy'n diddymu llawer o ddarpariaethau Ddeddf 1967 ac a sicrhaodd fod sefydliadau cyhoeddus sy'n gwasanaethu'r cyhoedd yng Nghymru, lle bynnag y maent wedi'u lleoli yn y DU, yn darparu gwasanaethau i'r cyhoedd yng Nghymru ar sail cydraddoldeb rhwng y Gymraeg a'r Saesneg. Sicrhawyd hyn gan 'gynlluniau iaith Gymraeg', dogfennau sydd i'w teilwra i amgylchiadau pob sefydliad unigol. Yn ogystal, sefydlodd y Ddeddf Fwrdd yr Iaith Gymraeg statudol i 'hyrwyddo a hwyluso'r defnyddio'r Gymraeg.' Cafodd y Bwrdd ei ddiddymu ym mis Mawrth 2012, gan y ddeddfwriaeth ddiweddaraf a ddisgrifir isod.
- Deddfau Llywodraeth Cymru (1998 a 2006), sy'n caniatáu datganoli pŵer mewn meysydd cyfyngedig i Gynulliad Cenedlaethol Cymru, ac yn rhoi iddo'r gallu i wneud 'unrhyw beth o fewn ei bŵer' i hyrwyddo'r Gymraeg.
- Ymhlith datblygiadau arwyddocaol eraill mae mudiadau cymdeithas sifil fel Mudiadau Dathlu'r Gymraeg (mudiad ymbarél o fudiadau Cymraeg), Dyfodol i'r Gymraeg, lobi o siaradwyr Cymraeg amlwg a'r Awr Gymraeg; yn ystod yr awr hon anogir siaradwyr Cymraeg, unwaith yr wythnos, i ddefnyddio'r Gymraeg ar Twitter bob wythnos.

- Arsyllfa polisi, a gyhoeddwyd ym mis Ionawr 2013 gan Gomisiynydd y Gymraeg, i astudio goblygiadau'r Gymraeg ym mhob maes polisi
- Y Coleg Cymraeg Cenedlaethol, sefydliad rhithwir lefel Prifysgol wedi'i ariannu gan y Llywodraeth, sy'n cydlynu darpariaeth Gymraeg yn sefydliadau addysg uwch Cymru. Mae darlithyddiaethau mewn technoleg wedi eu hysbysebu.

Mae llawer o'r datblygiadau uchod, i raddau, yn defnyddio neu'n creu angen am dechnoleg iaith.

3.4 HYRWYDDO A RHEOLEIDDIO IAITH

Un o'r datblygiadau mwyaf arwyddocaol ym mholisi iaith yng Nghymru fu caniatáu i Gynulliad Cenedlaethol Cymru Gymhwysedd Deddfwriaethol i greu deddfwriaeth mewn materion sy'n ymwneud â sawl agwedd ar y Gymraeg. Fe wnaeth hyn alluogi Llywodraeth Cymru i gyflwyno un o brif bileri ei Gytundeb Clymblaid 'Cymru'n Un', h.y. i ddrafftio Mesur y Gymraeg (Cymru) 2011. Mae'r Mesur yn rhoi i Weinidogion Cymru y pŵer i ddiddymu Bwrdd yr Iaith Gymraeg, ac i ad-drefnu ei swyddogaethau. Swydd Comisiynydd y Gymraeg oedd un o brif elfennau'r ddeddfwriaeth hon. Daeth y Comisiynydd i fodolaeth ym mis Ebrill 2012. Mae gan y Comisiynydd bŵer i orfodi cydymffurfiaeth gyfreithiol â chyfres newydd o 'Safonau Iaith Gymraeg', y bwriedir iddynt gymryd lle'r 541 Cynlluniau Iaith Gymraeg sydd ar waith, prif offeryn Deddf yr Iaith Gymraeg ar hyn o bryd. Mae'r safonau hyn yn cael eu rhannu fel a ganlyn:

- Safonau cyflenwi gwasanaethau
- Safonau llunio polisi
- Safonau gweithredu
- Safonau hybu
- Safonau cadw cofnodion

Mae'r safonau hyn wedi bod yn destun ymgynghoriad cyhoeddus gan y Comisiynydd ac yn awr maent wrthi'n cael eu hailddrafftio a'u harchwilio ymhellach gan Weinidogion Cymru. Disgwylir y bydd y drefn yn gwbl weithredol erbyn diwedd 2014. Mae'r safonau wedi eu hategu gan system cydymffurfio fwy trylwyr, gan gynnwys hysbysiadau cydymffurfio a dirwy o hyd at £5,000 (ac, wrth gwrs, y niwed i enw da a ddaw yn sgil hyn). Y tair safon gyntaf sydd fwyaf perthnasol i dechnoleg iaith. Maent yn sicrhau bod y gwasanaethau a ddarperir i'r cyhoedd yng Nghymru (drwy ba fodd bynnag) yn cael eu darparu trwy gyfrwng dewis iaith y defnyddiwr terfynol a bod yn rhaid i'r dewis hwnnw gael ei gipio a'i ailddefnyddio yn ei holl ymwneud â'r sefydliad hwnnw (gwefannau, ap, CRM ac ati).

> Gellir dwyn elfennau o'r sector teleffoni o fewn cylch gorchwyl Mesur y Gymraeg, a thrwy hynny ei gwneud yn orfodol i ryngwynebau teleffoni symudol i fod ar gael yn Gymraeg

Bydd safonau llunio polisi yn sicrhau bod yn rhaid i'r Gymraeg gael ei hystyried fel ffactor ym mhob penderfyniad polisi y bydd sefydliad yn ei wneud. E.e. o ran strategaeth tymor-hir TG sefydliad, 'a yw system X ar gael, neu'n debygol o fod ar gael yn y Gymraeg, ac a yw'n cynnal ac yn rheoli dewis iaith'—os nad yw'n gwneud, dylid ei ddiystyru, neu lunio achos dros barhau â'r broses gaffael (gan gadw cofnodion priodol o'r rhesymeg y tu ôl i'r penderfyniad i barhau). Mae'r safonau gweithredu yn ystyried gwaith mewnol sefydliad penodol, a byddai'r hawliau a allai godi o'r safonau hyn yn galluogi gweithwyr i gyfathrebu â'i gilydd yn Gymraeg heb osod rhwystr rhag hynny, a hynny â chefnogaeth statudol lawn. Mae technoleg yn hwylusydd allweddol ar gyfer hyn a disgwylir cynnig gweithredol, er enghraifft, o wasanaeth neu ddarpariaeth (o ran yr ochr weinyddol

a chynnwys) mewn rhaglenni o'r fath fel systemau rheoli cynnwys a rhyngwynebau TG. Mae'r Mesur hefyd yn caniatáu ar gyfer elfennau pwysig o'r sector telathrebu i gael eu dwyn o fewn ei gylch gorchwyl yn ddiweddarach a thrwy hynny ei gwneud yn ofynnol, er enghraifft, i ryngwynebau teleffoni symudol fod ar gael yn y Gymraeg.

Yn olaf fel cefndir strategol i'r fframwaith polisi sy'n effeithio ar dechnoleg iaith yng Nghymru, mae Llywodraeth Cymru wedi cyhoeddi dogfen Strategaeth Iaith sylweddol (a hynny wedi cyhoeddi dogfen Strategaeth ar gyfer Addysg Gymraeg yn barod) a gyhoeddwyd o dan y teitl *Iaith Fyw: Iaith Byw*. [16] Mae'r strategaeth yn amlinellu gweledigaeth y Llywodraeth ar gyfer Cymru ddwyieithog y mae'n dymuno ei gweld yn y dyfodol, ac mae Gweinidogion wedi datgan sawl gwaith yn gyhoeddus y bydd cynyddu'r defnydd o ddarpariaeth ieithyddol sydd ar gael yn un o brif sylfeini athronyddol y strategaeth honno. Mae'r athroniaeth hon yn cael ei hamlinellu, a'i harchwilio, yng nghyd-destun y ddarpariaeth TG sydd eisoes ar gael yn y Gymraeg, isod. Mae'r strategaeth yn rhagweld yn 'cynrychiolaeth gref o'r Gymraeg ar draws y cyfryngau digidol' ac yn neilltuo un o'i chwe phennod i 'Isadeiledd' ar gyfer yr iaith, lle yr eir i'r afael â thechnoleg iaith. Dyma dargedau strategol y bennod honno, gan nodi y bydd Llywodraeth Cymru yn gweithredu drwy:

- annog darparwyr gwasanaeth y sector preifat mawr, gan gynnwys banciau, manwerthwyr, cwmnïau ffôn symudol, datblygwyr meddalwedd a chaledwedd, ac eraill i ddatblygu gwasanaethau, rhaglenni a rhyngwynebau ar-lein drwy gyfrwng y Gymraeg
- hwyluso datblygu rhyngwynebau Cymraeg ar gyfer y cyfryngau rhwydweithio cymdeithasol a ddefnyddir yn gyffredin, gan gynnwys meddalwedd cod agored
- darparu, o bosibl ar sail gyfatebol, cyllid cychwynnol ar gyfer mentrau fel y rhain ar sail gynyddrannol dros gyfnod o amser
- datblygu consensws mewn meysydd blaenoriaeth lle mae angen buddsoddiad technolegol

Sefydlwyd y Gweithgor Gweinidogol yn gynnar yn 2012 a chyfarfu sawl gwaith i drafod Strategaeth ddrafft a Chynllun Gweithredu ar gyfer y Gymraeg a thechnoleg. Mae'r strategaeth [17] yn yn ymdrin yn fanwl â'r holl themâu uchod, ac yn rhoi pwyslais sydd i'w groesawu ar greu cynnwys, a fydd, wrth gwrs, yn asio'n gryf gyda'r offer a themâu technoleg iaith a ddisgrifir yn y ddogfen hon yn achos y Gymraeg. Mae'r Strategaeth, yn seiliedig ar drafodion y Gweithgor ac yn cynnig pum maes gweithredu:

- Marchnata a chodi ymwybyddiaeth
- Dylanwadu ar gwmnïau meddalwedd mawr
- Annog datblygu pecynnau meddalwedd newydd a gwasanaethau digidol drwy gyfrwng y Gymraeg
- Annog creu a rhannu, a defnyddio cynnwys digidol Cymraeg
- Rhannu arfer gorau yn y sectorau cyhoeddus, preifat a'r trydydd sector.

Bydd y rhain yn cael eu rhoi ar waith drwy gyllid, anogaeth a deddfwriaeth gan y Llywodraeth.

Er mwyn gwireddu gweledigaeth y Llywodraeth o Gymru ddwyieithog, rhaid i'r Gymraeg fod â lle haeddiannol ym myd technoleg, ac mae angen dull gweithredu strategol, hirdymor i sicrhau hyn. Fel y gwelwyd, mae hyn ar waith. Mae ystyriaethau eraill hefyd yn dylanwadu ar rôl a sefyllfa gyfreithiol y Gymraeg. O fewn y degawd diwethaf, daeth y Gymraeg yn rhan o'r agenda cydraddoldeb ond eto nid yw'n ymddangos yn amlwg o'i chymharu â meysydd cydraddoldeb eraill yn seiliedig ar hil, rhyw, cyfeiriadedd rhywiol neu anabledd. Un ffordd o gyfrannu at hollbresenoldeb y Gymraeg mewn technoleg oedd cyhoeddi canllawiau manwl, technegol a fyddai'n darparu cyngor ar sut i greu meddalwedd neu wefannau amlieithog, gan bwysleisio'r angen am newid

iaith yn hawdd. Ym mis Ebrill 2006, lansiodd Bwrdd yr Iaith Gymraeg *Ganllawiau a Safonau Meddalwedd Dwyieithog* cynhwysfawr, [18] (ar yr un diwrnod â'r ddogfen Strategaeth gyntaf ar gyfer TG a'r Gymraeg [19]) a'u dosbarthu i bob sefydliad a oedd â Chynllun Iaith statudol o dan Ddeddf yr Iaith Gymraeg 1993. Wedyn, cynhaliodd nifer o seminarau ar gyfer ymarferwyr technegol i roi cyhoeddusrwydd i'r safonau, un yr un yng ngogledd a de Cymru, ac un arall yn Llundain (yn bennaf ar gyfer Cyrff y Goron â Chynllun Iaith o dan Ddeddf yr Iaith Gymraeg 1993). Y gobaith oedd y byddai'r cyngor a roddwyd yn y ddogfen hon, a'r fframweithiau monitro a ddefnyddir i dracio sut y'i gweithredir, yn gwella darpariaeth ddwyieithog gwasanaethau electronig o bob math, gan wella ar y perfformiad a nodir yn nau *Giparolwg* o wefannau sector cyhoeddus y Bwrdd a gynhaliwyd yn 2001 [20] a 2003 [21]. Ym mis Awst 2009, lansiodd y Bwrdd Gynllun Achredu technegol [22] ar gyfer y ddogfen safonau ar ei wefan. Mae'r cynllun hwn, sy'n anelu at staff technegol medrus TG mewn sefydliadau â Chynllun Iaith o dan Ddeddf 1993 (ac unrhyw sefydliad arall sy'n dymuno darparu gwasanaethau TG dwyieithog) y Gymraeg, yn troi'r *Canllawiau a Safonau Meddalwedd Dwyieithog* i ffurf cwestiwn, gan alluogi'r staff technegol hynny i ganfod yn hawdd a yw system benodol yn cydymffurfio ac yn cyflenwi dewis iaith ar sail cydraddoldeb rhwng y Gymraeg a'r Saesneg. Y bwriad yw y byddai canlyniadau'r cwestiynau hyn yn cael eu defnyddio fel dangosyddion diacronig ar gyfer gwella gwasanaethau TG dwyieithog. Diweddarwyd y ddogfen hon, a'i hail-weithio fel un o'r dogfennau cyngor cyntaf a gyhoeddwyd gan Gomisiynydd y Gymraeg yn 2012. Bydd holl ddogfennau cyngor y Comisiynydd, maes o law, yn cael eu cyhoeddi fel Codau Ymarfer dan Fesur y Gymraeg (Cymru) 2011, gyda sefydliadau yn gorfod profi sut y maent wedi rhoi 'sylw dyledus' i'r codau hyn i gydymffurfio â'r drefn safonau a ddisgrifir uchod.

3.5 HYFFORDDIANT IAITH A THECHNOLEG YM MYD ADDYSG

Mae ar faes hyfforddiant defnyddiwr a meithrin gallu trwy gyfrwng y Gymraeg ym maes TG angen sylw penodol. Mae amryw o switiau hyfforddi yn bodoli, ar ffurf electronig ac ar bapur, yn eu plith y *Drwydded Yrru Gyfrifiadurol Ewropeaidd* [23] a phrosiect *Llythrennedd Digidol* Microsoft, sy'n defnyddio terminoleg safonol a chipluniau fersiynau Cymraeg ei gynhyrchion ei hun. Gan fod y byd TG yn symud mor gyflym, bydd angen diweddariadau i'r math hwn o hyfforddiant yn rheolaidd, ac mae pecynnau hyfforddi eraill wrthi'n cael eu paratoi. O ran hyfforddiant ar gyfer sefydliadau i ddarparu technoleg iaith, fe wnaeth Bwrdd yr Iaith Gymraeg gylchredeg Nodyn Cyngor o dan Ddeddf yr Iaith Gymraeg 1993, i bob un o'r 541 o sefydliadau a chanddynt gynllun iaith, gan egluro'r ddarpariaeth bresennol o ran technoleg iaith, gan nodi'r mythau a amlinellir yn y papur gwyn hwn—a'u chwalu. Mae modiwlau israddedig ar gael i wahanol raddau mewn gwahanol sefydliadau ond nid ymhob un.

3.6 AGWEDDAU RHYNGWLADOL

Mae'r Gymraeg wedi'i chwmpasu gan ddarpariaethau yn y Confensiwn Fframwaith ar gyfer Amddiffyn Lleiafrifoedd Cenedlaethol ac mae Llywodraeth y DU wedi cadarnhau 52 o gymalau Siarter Ewrop ar gyfer Ieithoedd Rhanbarthol neu Leiafrifol gyfer y Gymraeg. Mae hefyd wedi'i chynrychioli ar y Rhwydwaith ar gyfer Hyrwyddo Amrywiaeth Ieithyddol (y mae ei brif swyddfa yng Nghaerdydd).

3.7 Y GYMRAEG AR Y RHYNGRWYD

Mae ffigurau Llywodraeth Cymru [24] ar gyfer 2012 yn dangos bod gan 70% o gartrefi Cymru fynediad i'r rhyngrwyd. Mae hyn yn cyfateb i tua 77% o bobl 18 oed neu'n hŷn a chanddynt fynediad i'r rhyngrwyd yn y cartref. Dywedodd 73% o bobl eu bod yn defnyddio'r rhyngrwyd yn y cartref, gwaith neu mewn man arall; roedd hyn yn amrywio yn ôl oedran gyda chyfran llawer mwy o bobl o dan 45 oed yn defnyddio'r rhyngrwyd na'r rhai 45 oed a throsodd. Defnyddwyr y rhyngrwyd o dan 25 mlwydd oedd yn fwy tebygol (41%) na defnyddwyr y rhyngrwyd 65 oed a throsodd (8%) oed i fod wedi cael mynediad i'r rhyngrwyd yng nghartref rhywun arall yn ystod y tri mis diwethaf. Dengys ffigurau diweddarach [25] mai'r ganran o boblogaeth Cymru nad *oedd erioed* wedi defnyddio'r rhyngrwyd oedd 17.5% (o'i chymharu â 14% o boblogaeth y DU yn ei chyfanrwydd. Fodd bynnag, nid oes ffigurau manwl gywir ar gael ynghylch pa iaith y bydd pobl Cymru yn ei defnyddio i gael mynediad i'r rhyngrwyd. Y Gymraeg yw'r 65ain iaith fwyaf cryf ar Wikipedia, yn ôl cynnwys [26] (a chanddi *oddeutu* 49,000 erthygl), sy'n profi hyfywedd ac egni'r gymuned cod agored o lunwyr cynnwys gwirfoddol. Mae nifer o borwyr eisoes â rhyngwyneb ar gyfer y Gymraeg (Internet Explorer, Opera, Firefox), a sawl un arall, heb feddu ar ryngwynebau Cymraeg, ond yn caniatáu i *locale* y porwr gael ei newid i Gymraeg er mwyn awtomeiddio cyflwyno cynnwys Cymraeg. Yn anffodus, er mwyn manteisio ar y ddarpariaeth hon, rhaid gwybod beth yw *locale*, gwybod ymhle y mae, bod am ei ddefnyddio yn Gymraeg, a gwybod sut mae ei newid, a'i newid yn ôl os, er enghraifft, y bydd diweddariad meddalwedd yn ei ailosod. Mae hyn i gyd yn annibynnol ar iaith rhyngwyneb y system gweithredu er enghraifft, Microsoft Windows. Mae rhyngwyneb Cymraeg wedi cael ei ddatblygu ar gyfer Google (chwilio) a Gmail. Y cyfleuster rhyngrwyd a ddefnyddir amlaf yw chwilio'r we, ac mae'n hollbresennol ar bob math o ddyfais a phlatfform. Yn ogystal, mae chwilio ar y we ei hun yn defnyddio neu fe all ddefnyddio ystod o dechnolegau iaith (o wahanol lefelau o soffistigedigrwydd) i wella canlyniadau ac ansawdd yn gyffredinol. Ar wahân i'r bri o fod yn gysylltiedig â brand rhyngwladol mor llwyddiannus, mae rhyngwynebau Cymraeg Google yn darparu mwy na phrofiad rhyngrwyd addas ar gyfer siaradwyr Cymraeg, ond hefyd maent yn adlewyrchu'r angen cynyddol am offer chwilio a gwasanaethau prosesu iaith priodol i ddelio â data'r Gymraeg. Mae'r adran nesaf yn rhoi cyflwyniad i dechnoleg iaith a'i meysydd craidd, ynghyd â gwerthusiad o gefnogaeth gyfredol technoleg ar gyfer y Gymraeg.

4

CYMORTH TECHNOLEG IAITH AR GYFER Y GYMRAEG

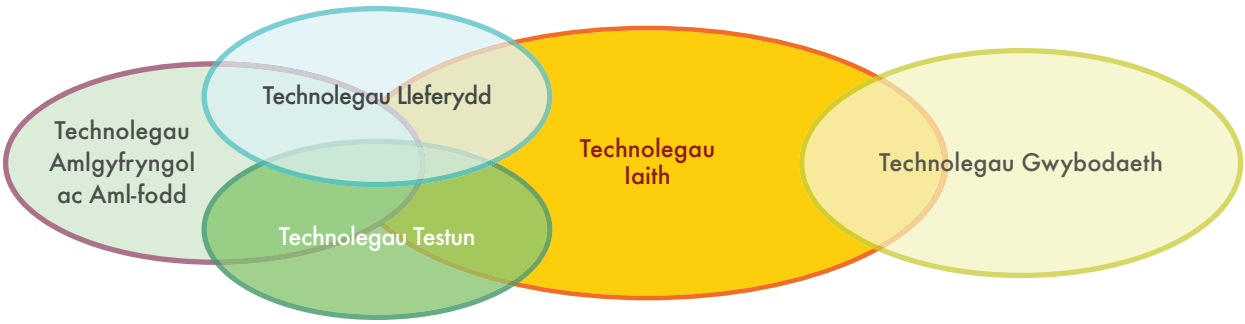
Defnyddir technolegau iaith i ddatblygu systemau meddalwedd sydd â'r nod o ymdrin ag iaith ddynol, ac felly yn aml fe'u gelwir yn 'dechnolegau iaith ddynol'. Daw iaith ddynol mewn ffurfiau llafar ac ysgrifenedig. Er mai lleferydd yw'r ffurf hynaf ar gyfathrebu ieithyddol o ran yr esblygiad dynol, bydd y rhan fwyaf o wybodaeth gymhleth yn cael ei storio a'i throsglwyddo drwy'r gair ysgrifenedig. Bydd technolegau testun a lleferydd yn prosesu neu yn cynhyrchu'r gwahanol fathau hyn o iaith, gan ddefnyddio geiriaduron, rheolau gramadegol, a semanteg. Mae hyn yn golygu bod technoleg iaith (TI) yn cysylltu iaith â gwahanol fathau o wybodaeth, yn annibynnol ar y cyfryngau (lleferydd neu destun) y mae'n cael ei fynegi ynddynt. Mae Ffigur 1 yn dangos y tirwedd technoleg iaith. Pan fyddwn yn cyfathrebu, rydym yn cyfuno iaith â dulliau eraill o wybodaeth a chyfryngau cyfathrebu—er enghraifft, gall siarad gynnwys ystumiau a mynegiant y wyneb. Bydd testunau digidol yn cysylltu â lluniau a seiniau. Gall ffilmiau gynnwys iaith mewn ffurf lafar ac ysgrifenedig. Mewn geiriau eraill, mae technolegau lleferydd a thestun yn gorgyffwrdd ac yn rhyngweithio â chyfathrebu amlfodd eraill a thechnolegau aml-gyfryngol. Yn yr adran hon, trafodir prif feysydd defnyddio technoleg iaith ar gyfer y Gymraeg, h.y., gwirio iaith, chwilio ar y we, rhyngweithio lleferydd, a chyfieithu awtomatig. Gall y rhaglenni a'r technolegau sylfaenol hyn gynnwys, ond nid ydynt yn gyfyngedig i:

- gywiro sillafu
- cymorth i greu cynnwys
- dysgu iaith gyda chymorth cyfrifiadur
- galw gwybodaeth yn ôl
- echdynnu gwybodaeth
- crynhoi testun
- ateb cwestiwn
- adnabod llais
- synthesis lleferydd

Mae technoleg iaith yn faes ymchwil sefydledig a chanddo set helaeth o lenyddiaeth ragarweiniol. Cyn trafod y meysydd uchod, ceir disgrifiad cryno o bensaernïaeth Rhaglen system TI nodweddiadol.

4.1 PENSAERNÏAETH RHAGLENNI

Fel arfer bydd meddalwedd ar gyfer prosesu iaith yn cynnwys sawl elfen sy'n adlewyrchu gwahanol agweddau ar iaith. Er bod rhaglenni o'r fath yn tueddu i fod yn gymhleth, dengys Ffigur 2 bensaernïaeth seml iawn system prosesu testun nodweddiadol. Mae'r tri modiwl cyntaf yn ymdrin â'r strwythur ac ystyr testun a roddwyd:



1: Technoleg iaith yn ei chyd-destun

1. Cyn-brosesu: glanhau'r data, dadansoddi neu gael gwared â fformatio, canfod pa iaith a fewnbynnwyd, ac yn y blaen.
2. Dadansoddiad gramadegol: dod o hyd i'r ferf, ei gwrthrychau, addaswyr ac elfennau brawddegol arall; canfod strwythur y frawddeg
3. Dadansoddiad Semantig: dadamwyso (h.y., canfod ystyr briodol geiriau mewn cyd-destun penodol); penderfynu anaffora (hy, pa ragenwau sy'n cyfeirio at ba enwau yn y frawddeg); cynrychioli ystyr y frawddeg mewn ffordd y gall peiriant ei darllen.

Wedi dadansoddi'r testun, gall modiwlau tasg-benodol berfformio gweithredoedd eraill, megis crynhoi awtomatig ac edrych mewn cronfa ddata. Noder bod pensaernïaeth y rhaglenni wedi'i symleiddio fawr ac wedi ei rhoi mewn ffordd 'ddelfrydol', er mwyn dangos cymhlethdod rhaglenni technoleg iaith mewn ffordd gyffredinol ddealladwy. Yng ngweddill yr adran hon, ceir gorolwg o gyflwr ymchwil ac addysg technoleg iaith ar gyfer y Gymraeg fel y mae heddiw, ynghyd â disgrifiad o ddatblygiadau technoleg iaith Gymraeg yn y gorffennol a'r presennol. Yn olaf, darperir amcangyfrif o offer ac adnoddau technoleg iaith greiddiol ar gyfer y Gymraeg o ran gwahanol ddimensiynau megis argaeledd, aeddfedrwydd ac ansawdd.

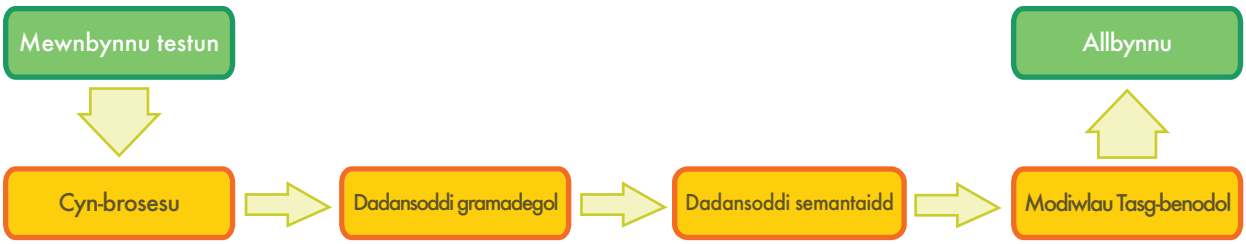
4.2 MEYSYDD RHAGLEN CRAIDD

Yn yr adran hon, canolbwyntir ar yr offer technoleg iaith pwysicaf a darperir gorolwgo weithgareddau TG ar gyfer y Gymraeg.

4.2.1 Gwirio Iaith

Bydd pawb sydd wedi defnyddio prosesydd geiriau megis Microsoft Word yn gwybod bod ganddo wirydd sillafu sy'n amlygu gallau sillafu ac yn cynnig cywiriadau. Ddeugain mlynedd wedi i Ralph Gorin greu'r rhaglen gyntaf i wirio sillafu, nid yw gwirwyr sillafu ond yn creu cymhariaeth seml o'r geiriau maent wedi eu hechdynnu, ond maent wedi dod yn gynyddol fwy soffistigedig. Gan ddefnyddio algorithmau ieithyddol ar gyfer dadansoddi gramadegol, maent yn canfod gwallau sy'n gysylltiedig a morffoleg (e.e. ffurfiau lluosog) yn ogystal â rhai sy'n deillio o'r gystrawen, megis berfau ar goll neu wrthdaro wrth gytuno rhwng berf-gorddrych (e.e., **she write a letter* yn Saesneg). Fodd bynnag, ni fydd y rhan fwyaf o wirwyr sillafu a gramadeg yn canfod gwallau yn y testun canlynol [27]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.



2: Pensaernïaeth nodweddiadol ar gyfer prosesu testun

Fel arfer, er mwyn mynd i'r afael â gwallau o'r fath, bydd angen dadansoddi'r cynnwys. Rhaid i'r math hwn o ddadansoddiad dynnu naill ai ar gramadegau sy'n benodol i'r iaith dan sylw, a'r rheini wedi eu codio'n llafurus i'r meddalwedd gan arbenigwyr, neu ar fodel ystadegol. Yn yr achos hwn, bydd model yn cyfrif y tebygolrwydd y bydd gair penodol yn ymddangos mewn safle penodol. Gellri creu model ieithyddol ystadegol yn awtomatig drwy ddefnyddio swm mawr o ddata ieithyddol (wedi'i hanodi). Gelwir hyn yn gorpws testun. Datblygwyd y ddau ddull uchod o gwmpas data'r Saesneg. Ar hyn o bryd, nid oes modd hawdd i drosglwyddo'r naill neu'r llall i'r Gymraeg oherwydd diffyg adnoddau ieithyddol sylfaenol. Ni cheir corpora testun digon mawr, a'r rheini wedi eu hanodi, i hyfforddi model ystadegol, ac ni chafwyd ymchwil ddigonol i amgodio gwybodaeth ieithyddol mewn gramadegau.

Nid yw gwirio iaith yn beth sydd ar gyfer prosesydd geiriau yn unig—mae hefyd yn berthnasol i systemau awduro

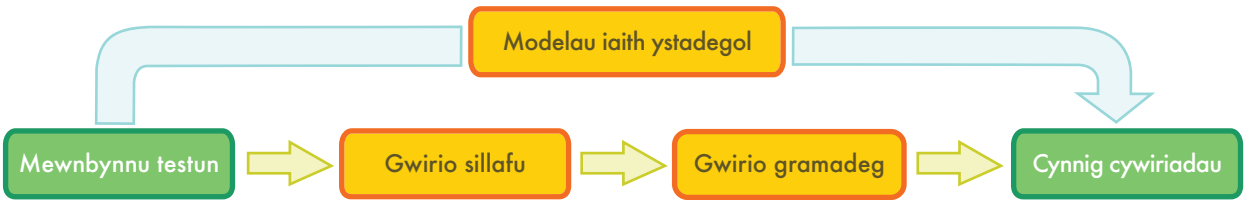
Y tu hwnt i'r gwirydd sillafu a gramadeg a chefnogaeth i awduron, mae gwirio iaith hefyd yn bwysig ym maes dysgu-iaith-â-chymorth cyfrifiadur. Dyma faes a fyddai o fydd sylweddol i ddysgwyr y Gymraeg ac ieithoedd bychain eraill.

4.2.2 Gwirwyr sillafu, Gwirwyr Gramadeg a Geiriaduron Cyfrifiadurol

Er mwyn mynd i'r afael â'r problemau o ran creu cynnwys a ddisgrifir uchod, mae nifer o wirwyr sillafu a gramadeg a geiriaduron cyfrifiadurol wedi eu comisiynu neu noddi gan gymorth grant. Y cyntaf oedd CySill, a gomisiynwyd gan Adrannau Seicoleg ac Ieithyddiaeth Prifysgol Cymru, Bangor. Roedd CySill yn chwyldroadol gan ei fod yn cywiro treigladau cytseiniaid cychwynnol y Gymraeg. Dilynwyd hyn gan Cysgair, geiriadur cyfrifiadurol sy'n atodi wrth raglenni prosesu geiriau, a grëwyd gan Ganolfan Bedwyr Prifysgol Cymru, Bangor. Fe wnaeth Canolfan Bedwyr ddiweddaru'r ddau yn 2004, a'u cynnwys ar un CD o adnoddau cyfrifiadurol Cymraeg, o dan yr enw Cysgliad. Ymysg y gwirwyr sillafu eraill sydd eisoes yn bodoli yn y Gymraeg ceir y canlynol:

- Gwirydd sillafu Cymraeg am ddim i Microsoft Office XP, 2003, 2007, 2010, 2013.
- Gwirydd sillafu Cymraeg am ddim i OpenOffice
- Gwirydd Sillafu Cymraeg ar gyfer Neo Office (Apple Mac)

Mae pwysigrwydd LAD (Language Autodetect), ffordd o adnabod ieithoedd mewn dogfen hyd yn oed yn fwy amlwg mewn lleoliadau dwyieithog. Sut, er enghraifft, byddai system adnabod llais yn adnabod a yw siaradwr newid yr iaith y maent yn ei siarad mewn cyfarfod? Yn



3: Gwirio iaith (uchaf: ystadegol; isaf: ar sail rheolau)

fwy syml, sut mae siaradwr Cymraeg yn llunio dogfennau dwyieithog Cymraeg/Saesneg ac yn eu prawfddarllen yn awtomatig gan y gwiriwr sillafu 'brodorol' heb fod y Gymraeg yn cael ei thagio'n anghywir fel petai'n Saesneg wedi'i chamsillafu? Mae Geiriadur yr Academi Gymreig (Saesneg-Cymraeg) [28] gynt ar bapur wedi'i lansio ar-lein, prosiect mawr a ariennir gan Fwrdd yr Iaith Gymraeg ac wedi hynny gan Gomisiynydd y Gymraeg.

4.2.3 Bysellfwrdd, marciau diacritig a ffontiau

Agwedd arall ar greu cynnwys yn y Gymraeg a oedd yn achosi anhawster sylweddol i ddefnyddwyr cyfrifiaduron yn y gorffennol oedd diffyg, yn y lle cyntaf, 'ŵ' ac 'ŷ' o'r set nodau safonol. I ddechrau, deliwyd â hyn drwy ddefnyddio ffontiau Cymraeg arbenigol sy'n dynwared y ffontiau system a gynhwyswyd gyda chyfrifiaduron adeg eu creu, ond a oedd yn cynnwys yr acen grom ar 'w' ac 'y'. Fodd bynnag, mae'r rhain yn achosi problemau wrth anfon ffeil i beiriant gwahanol nad yw'r ffontiau hyn wedi eu gosod arno, byddai'r nodau hyn yn cael eu disodli gan nodau Islandeg. Mae'r ffurf gywir o'r ddau nod hyn wedi eu cynnwys yn set nodau safonol Unicode (UTF-8) ers peth amser, gan osgoi'r angen ar gyfer prynu neu lawrlwytho ffontiau arbenigol ar gyfer defnyddwyr PC. Fodd bynnag, ni fu'n ddigon clir sut y dylai defnyddwyr gael mynediad i'r acenion hyn mewn ffordd safonol, gydag unigolion a sefydliadau gwahanol neu hyd yn oed yr un sefydliadau yn dewis llwybrau byr bysellfwrdd gwahanol, neu ddefnyddio rhifau cod i osod marciau diacritig i mewn i ffeiliau. Mae'r broblem hon o ddiffyg trawiadau bysell diacritig safonol bellach wedi'i datrys, ar gyfer defnyddwyr PC, mewn dwy ffordd: (1) gan Sgema Bysellfwrdd Estynedig Microsoft ar gyfer y Deyrnas Unedig, a (2) gan gynnyrch poblogaidd rhad ac am ddim o'r enw 'To Bach' (To) a grëwyd gan Dechnoleg Draig.

4.2.4 Technoleg Lleferydd a'r Gymraeg

Gall technoleg llais gynnwys cynhyrchu llais synthetig neu adnabyddiaeth o lais dynol gan system TG. Mae technoleg o'r fath eisoes yn dechrau treiddio i'n bywydau bob dydd (mae sawl canolfan galw wedi awtomeiddio llawer o'r prosesau maent yn eu defnyddio, mae rhai ffonau symudol sy'n gallu derbyn e-bost eisoes yn cynnig cyfleuster llais synthetig i ddarllen negeseuon e-bost yn uchel i'r derbynnydd). Gall technoleg lleferydd fod yn gaffaeliad i unrhyw raglen TG benodol. Gall symleiddio mynediad i ddata, cyflymu mewnbynnu data a chaniatáu rheolaeth ddi-ddwylo ac, yn arwyddocaol, darparu dilysu biometrig goddefol seiliedig-ar-lais ar gyfer mynediad i wasanaethau diogel megis bancio. Mae'n amlwg hefyd fod iddi ganlyniadau enfawr i'r rhai sydd â nam ar eu golwg. Gan fod angen amser sylweddol i amgodio corpws lleferydd i unrhyw iaith, yn hanesyddol, yr ieithoedd mwy sydd wedi elwa ar y buddsoddiad mwyaf, gyda'r ieithoedd lleiafrifol yn tueddu i gael eu gadael ar ôl. Gan fod arwyddion yn dangos y bydd technoleg lleferydd drwy gydgyfeiriant â rhag-

lenni bob dydd, yn dod yn rhan fwyfwy pwysig o fywyd bob dydd yn y dyfodol, mae'n bwysig bod y Gymraeg yn sicrhau troedle cryf yn y maes hwn. O sylwi ar bwysigrwydd strategol y maes, fe wnaeth Bwrdd yr Iaith Gymraeg gyd-ariannu, gydag INTERREG, brosiect WISPR [29] (Welsh and Irish Speech Processing Resources) yng Nghanolfan Bedwyr ym Mhrifysgol Bangor. Wedi hynny, esblygodd y prosiect hwn yn Brosiect SALT [30]. Yn gynnar yn 2010 cafwyd lleisiau o ansawdd uwch, yn seiliedig ar y lleisiau gwreiddiol sylfaenol. Creodd prosiect cychwynnol WISPR beiriant SAPI sylfaenol (Speech Application Programming Interface) ar gyfer y Gymraeg. Ymhlith llawer o ffyrdd eraill, gall peiriannau synthesis lleferydd gael eu defnyddio i drosi geiriau o ddogfen gyfrifiadurol (e.e. dogfen prosesydd geiriau, tudalen gwe), neu ryngwyneb yn lleferydd clywadwy i'w glywed drwy system sain y cyfrifiadur. Byddai hyn yn ddefnyddiol i bobl sydd angen neu eisiau dilysu llafar o'r hyn y maent yn ei weld mewn print. Wrth gynnal Eisteddfod Genedlaethol Cymru ym mis Awst 2010, lansiwyd fersiwn alffa o ddau lais synthetig (un gwrywaidd, un benywaidd) o ansawdd uchel. A hwythau wedi'u hariannu'n rhannol gan Lywodraeth Cymru, a hefyd gan Sefydliad Cenedlaethol Brenhinol y Deillion (RNIB), fe gafodd cynulleidfa'r lansiad eu symud gan ansawdd uchel y lleisiau, a ddarllenodd ddarn o nofel. Mae'r ffaith mai fersiwn alffa a oedd yn cael ei lansio yn golygu y bydd tonnau dilynol o wella (yn seiliedig ar adborth torfol drwy ryngwyneb gwe).

4.2.5 Adnabod llais

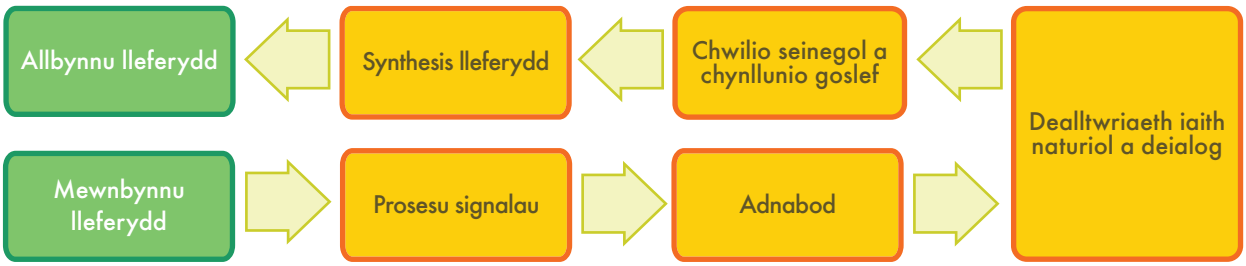
Mae a wnelo adnabod llais, (fel y'i diffinnir gan dîm Prosiect WISPR) neu 'lleferydd-i-destun', â dal a digideiddio tonnau sain o feicroffon, eu trosi i unedau iaith sylfaenol neu ffonemau, creu geiriau o'r ffonemau, a dadansoddi cyd-destun y geiriau i sicrhau sillafiad cywir am eiriau sy'n swnio'n fel ei gilydd (megis dear a deer yn Saesneg).Yna bydd y cynnyrch yn cael ei ddangos ar y sgrin fel testun. Unwaith eto, er y gall y manylion technegol ymddangos y tu hwnt i'r person lleyg, ni ddylid tanbrisio cyrhaeddiad ac arwyddocâd y cyfleuster hwn, gan fod llawer o systemau gweithredu cyfrifiadurol a setiau llaw *eisoes* yn cynnig cyfleusterau adnabod llais allan-o'r-bocs. Mae hyn yn debygol o gynyddu gyda threigl amser. Nid yw'n amhosib dychmygu y byddwn yn archebu bwyd, yn gwneud ein bancio a llu o wasanaethau eraill drwy gyfrwng adnabod llais yn y dyfodol. Mae ap Google Translate eisoes yn caniatáu i'r defnyddiwr ddefnyddio adnabod lleferydd, yn gysylltiedig â synthesis lleferydd trwy gyfrwng cyfieithu awtomatig. Gan gydnabod pwysigrwydd datblygu'r maes hwn, hefyd rhoddodd y Bwrdd grant i greu peiriant Adnabod Llais sylfaenol. Mae cydrannau'r prosiect hwn, a phrosiect synthesis lleferydd WISPR, ar gael am ddim ar-lein.

Ym mis Mai 2010, cyhoeddodd Panel Adolygu Annibynnol o Wasanaethau Dwyieithog Cynulliad Cenedlaethol Cymru ei adroddiad terfynol [31]. Ymhlith ei argymhellion lu, bu technoleg yn flaenllaw, gydag argymhelliad y dylid datblygu mwy o dechnoleg lleferydd er mwyn, yn y tymor hir, creu trawsgrifiadau o gyfarfodydd yn lled awtomatig. Nodwyd hefyd ddymuniad y Cynulliad i fod yn arweinydd yn y maes o ran darpariaeth ddwyieithog, drwy ddefnyddio technoleg.

Nid yw Mesur y Gymraeg (Cymru) 2011 yn gymwys i Gynulliad Cenedlaethol Cymru a Chomisiwn y Cynulliad (y Corff Corfforaethol sy'n gyfrifol am ddarparu eiddo, staff a gwasanaethau i gefnogi Aelodau'r Cynulliad).

Cafodd Bil Cynulliad Cenedlaethol Cymru (Ieithoedd Swyddogol) (Cymru) ei gymeradwyo gan y Cynulliad ar 3 Hydref 2012. Prif nod y Ddeddf yw gosod dyletswydd statudol ar Gynulliad Cenedlaethol Cymru a Chomisiwn y Cynulliad i gydnabod y Gymraeg a'r Saesneg yn ieithoedd swyddogol ac i drin y ddwy iaith yn gyfartal.

Mae'r Ddeddf hefyd yn gosod dyletswydd ar y Com-



4: System deialog yn seiliedig ar leferydd

isiwn i fabwysiadu a chyhoeddi Cynllun Ieithoedd Swyddogol [32] sy'n nodi'r camau y bydd yn eu cymryd i gydymffurfio â'i ddyletswyddau fel yr amlinellir yn y Ddeddf.

Cymeradwywyd y Cynllun gan y Cynulliad ar 17 Gorffennaf 2013. Mae'n diffinio'r safonau a'r gwasanaethau y gall Aelodau a'r cyhoedd eu disgwyl gan Gomisiwn y Cynulliad ac yn nodi pedwar maes allweddol ar gyfer gweithredu.

Bydd Comisiwn y Cynulliad yn:

- darparu cymorth arloesol, wedi'i deilwra, i alluogi pobl i ddefnyddio'r ddwy iaith yng nghyd-destun busnes y Cynulliad;
- buddsoddi'n sylweddol mewn technoleg fel ffordd o drawsnewid y ddarpariaeth o wasanaethau dwyieithog tra'n darparu gwerth am arian;
- datblygu sgiliau a hyder staff y Comisiwn i ddefnyddio'u Cymraeg; a
- rhannu profiadau o ddarparu gwasanaethau dwyieithog gyda sefydliadau eraill yng Nghymru a deddfwrfeydd eraill a cheisio dysgu ganddynt.

Ar hyn o bryd, mae defnydd y Cynulliad o dechnoleg cyfieithu yn gyfyngedig. Mae'r Google Translator Toolkit yn cael ei ddefnyddio gan gontractwyr allanol y Cynulliad i gyfieithu Cofnod y Trafodion ac ategir hyn gan waith golygu a phrawfddarllen i gywiro a mireinio allbwn y peiriant. Yn fewnol, mae Gwasanaeth Cyfieithu a Chofnodi'r Cynulliad yn defnyddio meddalwedd cof cyfieithu Wordfast, sydd wedi arwain at gynnydd mewn allbwn.

Mae'r Cynllun Ieithoedd Swyddogol yn ymrwymo'r Cynulliad i wneud y defnydd gorau o dechnoleg i gyfieithu dogfennau yn gyflymach ac yn fwy effeithlon. Mae wedi dechrau gwneud gwaith i archwilio potensial a manteision buddsoddi mewn system cyfieithu peiriant pwrpasol. Fel rhan o'r gwaith hwn, maent yn ystyried sut y gall cyfieithu awtomatig gael ei ddefnyddio nid yn unig gan y Gwasanaeth Cyfieithu a Chofnodi, ond gan aelodau eraill o staff y Cynulliad, staff cymorth Aelodau'r Cynulliad ac Aelodau'r Cynulliad, a'r posibilrwydd o'i wneud ar gael i sefydliadau y tu hwnt i'r Cynulliad.

4.2.6 Integreiddio Cyfieithu awtomatig a Thechnoleg Lleferydd

Un maes sy'n haeddu ystyriaeth yn y tymor canolig yw integreiddio technoleg llais gyda'r dechnoleg cyfieithu awtomatig a ddisgrifiwyd uchod. Byddai'r sefyllfa ddelfrydol yn galluogi dau o bobl, y naill yn siarad Cymraeg, a'r llall yn siarad Saesneg i sgwrsio â'i gilydd. Byddai hyn yn cael ei gyflawni drwy adnabod llais, bwydo i mewn peiriant cyfieithu awtomatig, ac allbynnu'r cyfieithiad perthnasol drwy gyfrwng synthesis lleferydd. Mae technolegau integreiddiol o'r fath eisoes yn cael eu cynhyrchu mewn sawl sefydliad, er enghraifft, yr Athro

Alex Waibel o Brifysgolion Carnegie Mellon a Karlsruhe yn benodol. Mae'n bosibl y bydd technoleg wedi'i hawtomeiddio o'r fath wedi'i chynnwys yn y systemau gweithredu y byddwn yn eu defnyddio bob dydd. Mae angen, felly, i ieithoedd llai fel y Gymraeg gael eu cynnwys yn y datblygiadau hyn. Mae potensial integreiddio o'r fath ar gyfer gwasanaethau Cymraeg, a chyfarfodydd dwyieithog, yn amlwg, a hynny mewn oes a reolir fwyfwy gan TG.

4.2.7 Cyfieithu a Therminoleg

Mae Comisiynydd y Gymraeg yn gyfrifol am bolisi cyfieithu yng Nghymru ac adeg llunio'r papur gwyn hwn roedd wrthi'n cynnal adolygiad annibynnol o'r proffesiwn a'r anghenion strategol. Pwrpas yr adran hon, felly, yw delio â *thechnoleg* iaith a chyfieithu, a'r cyfraniad y gall technoleg ei wneud i ddiwydiant cyfieithu'r Gymraeg neu unrhyw iaith arall a chanddi sefyllfa sosioieithyddol debyg.

4.2.8 Cyfieithu â Chymorth Cyfrifiadur

Dyma'r term ehangaf a ddefnyddir i ddisgrifio maes rhaglenni technoleg iaith sy'n awtomeiddio neu'n cynorthwyo'r weithred o gyfieithu testun o'r naill iaith i'r llall. Maent yn effeithiol iawn o ran gwella cynhyrchedd cyfieithu, yn enwedig o ran hwyluso cyfieithu cyflym iawn o destunau gwreiddiol ailadroddus. Dyma rai o'r mathau mwyaf cyffredin o gof cyfieithu: SDLX Trados, Déjà Vu, Wordfast ac yn fwyaf diweddar, y Google Translator Toolkit, Pootle, Transifex a OmegaT. Mae meddalwedd cof cyfieithu yn cael ei ddefnyddio eisoes gan ystod o sefydliadau cyhoeddus yng Nghymru a thu hwnt, yn eu plith, Llywodraeth Cymru, Prifysgol Caerdydd (drwy model torfoli) a Chynulliad Cenedlaethol Cymru ei hun. Un o brif rinweddau technolegau o'r fath yw y gall cyfieithwyr rannu gwaith cyfieithu ei gilydd, waeth beth fo'u lleoliad daearyddol, drwy gronfeydd data canolog o gofion cyfieithu a wasanaethir i gyfrifiaduron unigol drwy rwydweithiau corfforaethol neu hyd yn oed y rhyngrwyd. Mae hyn yn golygu y gall cyfatebiaethau 100% o segmentau'r testun gwreiddiol gael eu hailddefnyddio, gan arwain at gynnydd yng nghysondeb a chyflymder cyfieithu (am wybodaeth ar gysondeb terminoleg, gweler isod). Gall rhaglenni cof cyfieithu hefyd gynnig cyfieithiadau o gyfatebiaethau rhannol ('fuzzy') a geir yn eu cronfeydd data. Mae hyn hefyd yn cynyddu cyflymder y cyfieithu. Prif gafeat y dull hwn, wrth gwrs, yw bod ansawdd y cyfieithiadau a rennir drwy gofion cyfieithu a storir yn ganolog yn dibynnu ar ansawdd a chysondeb yr holl gyfieithiadau hynny a fwydir i mewn iddynt. Felly, yn amlwg mae angen rheolaeth olygyddol lefel uchel ar brosiectau ar raddfa fawr o'r fath. Un peth y dylid ei bwysleisio'n gyson [33] yw bod mawr angen cyfieithwyr sy'n fodau dynol, a nod y dechnoleg a ddisgrifir yma yw cynyddu eu cynhyrchedd a'u cysondeb—ac nid i'w disodli!

4.2.9 Rheoli Terminoleg

Unwaith eto, dylid nodi ar ddechrau'r adran hon nad lle'r papur cyfredol yw delio â maes safoni termau *fel y cyfryw*, dim ond i hwyluso lledaenu terminoleg safonol drwy TG fel un o amcanion y broses normaleiddio iaith. Mae technoleg yn bodoli eisoes, fel yn achos y meddalwedd cof cyfieithu a ddisgrifir uchod, i drosglwyddo rhestri safonol o derminoleg i ddefnyddwyr terfynol yn awtomatig. Mae hyn yn bwydo i mewn i raglenni rheoli terminoleg sy'n gysylltiedig â rhai rhaglenni cof cyfieithu. Mae rhai rhaglenni o'r fath hefyd yn cynnig cyfleusterau dadamwyso sy'n darparu gwybodaeth ychwanegol i ddefnyddwyr, e.e. mae'n bosibl na fydd fersiwn Gymraeg Mill Street yn Aberystwyth (Dan Dre), yr un cyfieithiad â Mill Street mewn unrhyw dref arall yng Nghymru (mae 'Stryd y Felin', wrth reswm, yn gyfieithiad gramadegol a chywir, ond ni chaiff ei ddefnyddio ar gyfer enw'r stryd yn Aberystwyth). Mae gan y gair 'access' lawer o ystyron, gan gynnwys fel enw sy'n disgrifio'r

lleoliad lle mae eir i mewn i adeilad, yn ogystal â berf sy'n golygu'r camau gwirioneddol a gymerir i gael mynediad i'r adeilad, ac yn wir cael mynediad i wybodaeth. Gall 'Mole' fod yn anifail tanddaearol bach sy'n effro liw nos, yn berson sy'n rhyddhau gwybodaeth o sefydliad, neu nifer eithriadol o fawr a ddefnyddir fel arfer ym myd Cemeg; gallai'r holl ystyron hyn gael eu gwahaniaethu drwy gyfleusterau dadamwyso. Po fwyaf maint, ansawdd a chysondeb cof cyfieithu, uchaf yw'r tebygolrwydd y bydd llwyth gwaith y cyfieithydd yn cael ei gyflawni mewn cyfnod byrrach. O feddu ar ymwybyddiaeth o'r posibiliadau y gall meddalwedd TM eu cynnig i ieithoedd lleiafrifol, yn gynnar yn 2010. Rhyddhaodd gyfanswm o tua hanner miliwn o eiriau o destun dwyieithog, wedi'u prawfddarllen a hynny ar ffurf TMX (safon y diwydiant ar gyfer cyfnewid cofion yn agored) gan gynnwys:

- Cof Cyfieithu Adnoddau Dynol
- Cof Cyfieithu Bwydlenni
- Cof cyfieithu a grëwyd drwy alinio gwefan Bwrdd yr Iaith Gymraeg (sy'n cynnwys llawer o eirfa'r sector cyhoeddus)

Fe wnaeth y datganiad i'r wasg a gyhoeddwyd ynghylch y datblygiad hwn alw ar y sector cyhoeddus cyfan i rannu eu cyfieithiadau yn agored, gyda'r Bwrdd yn cynnig i fod yn frocer ar gyfer y data mewn cyfnewidfa cof cyfieithu. Y prif wahaniaeth rhwng hyn a chyfnewidfeydd eraill yw bod y data a fyddai ar gael yn rhad ac am ddim. Mae'r holl gofion hefyd wedi eu llwytho i'r Google Translator Toolkit a ddisgrifir isod, ac felly yn ychwanegu at y corpws o gyfieithiadau sy'n seiliedig ar enghraifft ym mheiriant cyfieithu awtomatig Google, Google Translate. Mae'r weledigaeth hon wedi ysgyried datblygiadau ehangach ym myd gwe2.0, torfoli a dulliau cydweithredol. Yn ei hanfod, mae'n credu mai dim ond daioni a ddaw o rannu data o ansawdd rhwng sefydliadau. Enghraifft gorsyml a roddir yn aml mewn cyflwyniadau yw achos damcaniaethol 22 awdurdod lleol Cymru yn cyfieithu ffurflen Treth y Cyngor 22 o weithiau, tra gallai gweinyddion cof cyfieithu helpu i awtomeiddio hwn yn sylweddol. Mae'r fath rannu a ffiniau niwlog rhwng sefydliadau hefyd yn adlewyrchu tueddiadau ehangach mewn theori rheoli ac agendâu nifer o adroddiadau a gomisiynwyd y Llywodraeth, fel Adolygiad Gershon [34], Adolygiad Beecham, [35] Cod Ymarfer Llywodraeth y DU ar God Agored [36]. Mae'n hollbwysig sôn am reoli ansawdd y data a ryddheir gan sefydliadau cyhoeddus; un elfen o hyn fyddai cyfrinachedd. Gall cofion cyfieithu a ddefnyddir mewn sefydliadau cyhoeddus gynnwys data personol yn ymwneud ag unigolion y gellir eu hadnabod y mae'n rhaid ei symud cyn ei gyhoeddi er mwyn osgoi torri Deddf Diogelu Data'r DU. Mae dulliau o awtomeiddio hwn eisoes ar gael mewn systemau o'r fath. Ni ddylai technoleg o unrhyw fath fod yn ddiben ynddo'i hun ond yn alluogwr polisi strategol, yn achos y Gymraeg, ar gyfer normaleiddio ieithyddol.

4.2.10 Llif gwaith cyfieithu a Rheoli Dogfennau Dwyieithog

Gellir cynorthwyo rheoli llawer o brosiectau cyfieithu cydamserol drwy TG a systemau llif gwaith cyfrifiadurol. Gall y rhain hefyd fonitro argaeledd gwahanol gyfieithwyr allanol neu lawrydd i ymgymryd â gwaith ychwanegol (megis nifer o eiriau y gellir eu cyfieithu'r dydd fesul thema, a'r gyfradd cost). Gall systemau 'dashfwrdd' mwy soffistigedig o'r fath hyd yn oed ryngwynebu ag offer CAT er mwyn lleihau cyfieithu ailadroddus sy'n cael ei gontractio allan. Gall technoleg o'r fath hefyd, wrth gwrs, reoli swyddfeydd, er enghraifft, mewn sefydliadau megis awdurdodau lleol sydd â phyllau o gyfieithwyr mewnol. Byddai hyd yn oed yn bosibl i gyfieithwyr llawrydd allanol gofrestru ar gyfer gwasanaethau o'r fath er mwyn cael gwaith rheolaidd, a diweddariadau i gof cyfieithu a therminoleg gyson, a thrwy hynny greu corpws o gofion cyfieithu sy'n cyson

esblygu a gwella y gellir ei rannu ymhellach, gan gyfrannu ymhellach at normaleiddio'r Gymraeg. Mae systemau TM yn y 'cwmwl' megis y Google Translator Toolkit, yn hwyluso cydweithio o'r fath ymhellach o ran rheoli terminoleg a chof cyfieithu.

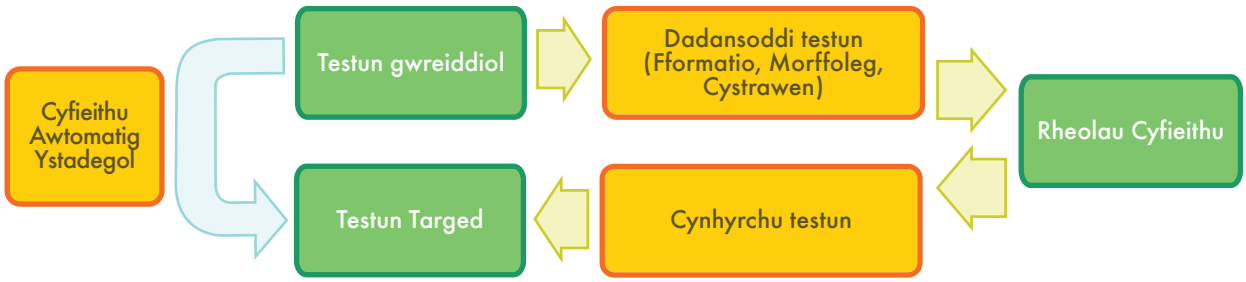
4.2.11 Llif gwaith cyfieithu a Systemau Rheoli Cynnwys

Mae'r systemau cof cyfieithu a ddisgrifir uchod yn gallu cael eu plygio i mewn i systemau rheoli cynnwys/dogfennau o bob math, a'u cysylltu â rhaglenni llif gwaith ar gyfer unedau/cwmnïau/gweithwyr cyfieithu llawrydd ac i systemau cyfieithu awtomatig. Mewn byd lle mae gwybodaeth yn cyson newid, mae'n bwysig bod y ddau fersiwn ieithyddol yn cael eu cyhoeddi a'u diweddaru ar yr un pryd. Yn y modd hwn, dylid cofio mai prif egwyddor Mesur y Gymraeg (Cymru) 2011 yw na ddylai'r Gymraeg gael ei thrin yn llai ffafriol na'r Saesneg. Mae gan dechnoleg rôl ganolog i'w chwarae o ran galluogi rheoli'r fath gyhoeddi cydamserol. Er enghraifft, ar wefan awdurdod llywodraethol neu leol sy'n gwasanaethu'r cyhoedd yng Nghymru, bydd rhannau o dudalennau neu baragraffau yn cael eu diwygio bob dydd. Pan fydd y system rheoli cynnwys yn cael ei rheoli gan siaradwyr di-Gymraeg, gall y dechnoleg hon gyfeirio segmentau i gyfieithydd a fydd wedyn yn eu cyfieithu gan ddefnyddio cof cyfieithu (a meddalwedd rheoli terminoleg). Pan fydd y cyfieithydd wedi cwblhau cyfieithu, bydd y system llif gwaith yn llwybro'r cyfieithiad yn ôl i'r system rheoli cynnwys yn awtomatig a fydd wedyn yn ei gyhoeddi ar yr un pryd â'r fersiwn Saesneg gwreiddiol a storiwyd wrth aros am ei gymar Cymraeg. Mae hwn yn offeryn pwysig ar gyfer normaleiddio'r pellach ar y Gymraeg a fydd yn galluogi defnyddwyr i gael mynediad at testun Cymraeg a Saesneg cyfredol. (Un o ganlyniadau'r *Ciparolygon* o wefannau oedd nad oedd fersiynau Cymraeg rhai o wefannau'r sector cyhoeddus yn cael eu diweddaru mor gyson â'r fersiynau Saesneg, gwelid 'Fersiwn Cymraeg i ddilyn', mewn rhai achosion, am sawl blwyddyn).

4.2.12 Cyfieithu Awtomatig

Mae cyfieithu awtomatig yn thema a godir yn aml wrth drafod y Gymraeg, cyfieithu a materion TG. Mae wedi ymelwa ar nifer o flynyddoedd o waith ymchwil, ac mae rhai datblygiadau mawr wedi eu sicrhau. Fodd bynnag, ni all technoleg eto gynnig gwaith cyfieithu o ansawdd i gyd-fynd â gwaith cyfieithwyr dynol heb ôl-olygu gan fodau dynol. Serch hynny, mae'r defnydd o systemau arlein rhad ac am ddim, yn enwedig Google Translate, a ryddhawyd ar gyfer y Gymraeg ddiwedd Awst 2009, yn dangos bod modd (yn achos mathau penodol o destun) darparu lefel ddefnyddiol a defnyddiadwy o gyfieithu. Hefyd, mewn cyfuniad â rheolaethau ar yr iaith a ddefnyddir er enghraifft mewn ysgrifennu technegol, gall MT ddarparu cyfieithiad drafft cyntaf o ansawdd ardderchog nad oes angen fawr ddim diwygio arno ac sy'n cynnig arbedion mawr o ran costau cyfieithu. Gall hefyd ddarparu bras gyfieithiad o'r Gymraeg, sy'n galluogi mynediad i'r gymuned ryngwladol, a galluogi staff di-Gymraeg i ddelio â gohebiaeth ysgrifenedig gan gydweithwyr/aelodau o'r cyhoedd sy'n siarad Cymraeg.

Fe argymhellodd astudiaeth ddichonolrwydd Bwrdd yr Iaith Gymraeg 2004 i gyfieithu awtomatig gan yr Athro Harold Somers o Brifysgol Manceinion, a Golygydd yr *International Journal of Machine Translation*, greu peiriant cyfieithu hybrid ag iddo dair rhan: EBMT [Cyfieithu Peirianyddol yn seiliedig ar Enghreifftiau], RBMT [Cyfieithu Peirianyddol yn seiliedig ar Reolau], ac SMT [Cyfieithu Peirianyddol ar Sail Ystadegau]. Byddai hyn yn galluogi bras gyfieithu rhwng y Gymraeg a'r Saesneg, ac, yn y dyfodol, yn integreiddio â rhaglenni eraill (megis cwarel ymchwil MS Office a systemau cof cyfieithu). Fel y nodwyd uchod, dylid ystyried hyn yn gymorth i gyfieithwyr, golygyddion ac eraill, yn hytrach nag yn fodd i'w disodli [37]. Byddai symud o



5: Cyfieithu awtomatig (chwith: ystadegol; dde: ar sail rheolau)

gyfieithu, i *ôl-olygu*, fodd bynnag, yn newid diwylliannol sylweddol ar gyfer nifer o gyfieithwyr a dylid defnyddio methodoleg rheoli newid cydnabyddedig lle mae system o'r fath yn cael eu rhoi ar waith. O ran hynny, dylid nodi bod Google Translate—System Cyfieithu Awtomatig Google, yn cysylltu â'r Google Translator Toolkit a hefyd SDL Trados Studio 2009, OmegaT a systemau cof cyfieithu eraill. Mae hyn yn rhoi i'r cyfieithydd proffesiynol ddashfwrdd llawn o opsiynau cyfieithu 'cwmwl': cyfieithu awtomatig [sy'n seiliedig ar enghraifft], Cof Cyfieithu, a rhestrau terminolegol, i gyd ag iddynt fantais o rannu TM mewn amser go iawn gyda'r holl ddefnyddwyr eraill. Mae hyn uwchlaw y grŵp bach o beiriannau cyfieithu awtomatig megis Intertran sydd wedi eu seilio ar ystadegau; mae'r rhain, o'u defnyddio gan amaturiaid a chanddynt ewyllys da ond heb fod ganddynt gymwysterau wedi creu rhai o'r cyfieithiadau mwyaf rhyfedd a welwyd erioed, e.e. cyfieithwyd 'Staff Entrance' 'yn gywir' fel 'Pastwn Taflu i Berlewyg' [hy [darn mawr o bren [a 'staff'] /taflu rhywun i mewn i berlewyg [to 'entrance' them]]. Mae enghreifftiau eraill yn rhy niferus i'w crybwyll a byddant, yn ddiau, yn gyfarwydd i ddarllenwyr mewn ardaloedd dwyieithog eraill. Mae llawer o enghreifftiau i'w gweld ar y wefan rhannu lluniau *Scymraeg* [38]. Mae'r peiriannau cyfieithu awtomatig canlynol ar gael ar gyfer y Gymraeg:

- Apertium, a ddatblygwyd gan y grŵp ymchwil Transducens yn y departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant mewn cydweithrediad â Pheirianneg Iaith Prompsit.
- Google Translate
- System Cyfieithu awtomatig Saesneg-Cymraeg Alffa Prifysgol Bangor
- Intertran

Gallai Awtomeiddio Cyfieithu gynorthwyo ieithoedd lleiafrifol drwy wella cynhyrchedd cyfieithwyr ieithoedd lleiafrifol a thrwy rannu adnoddau iaith ymhlith aelodau o'r gymuned cyfieithu proffesiynol. Gall cynnydd cynhyrchedd cyfieithu ynghyd â chynnydd o ran ansawdd cyfieithu arwain at ostyngiad mewn costau cyfieithu i iaith leiafrifol, er nad dyma'r unig reswm dros bledio achosion mabwysiadu technoleg o'r fath. Gallai cynnydd yng nghyflymder y cyfieithu (drwy, er enghraifft, lleihau cyfieithu ailadroddus) arwain at fwy o gyfieithiadau'n cael eu gwneud. Os bydd mwy o gyfieithiadau iaith leiafrifol yn cael eu gwneud mae mwy o gyfle y bydd mwy o gynnwys ar gael yn yr iaith, sy'n cynorthwyo i 'normaleiddio' iaith. Mae gwneud ieithoedd lleiafrifol yn fwy gweladwy, yn enwedig mewn technolegau modern, yn debygol o godi statws y ieithoedd lleiafrifol yn y llygaid siaradwyr iaith leiafrifol ac o bosibl eu dymuniad, a'u cyfle, i ddefnyddio eu ieithoedd lleiafrifol. Y Google Translator Toolkit, adeg ysgrifennu'r

papur gwyn hwn, oedd yr unig raglen TM a oedd ar gael am ddim a honno'n seiliedig ar dechnoleg cwmwl heb fod angen unrhyw wybodaeth am beirianneg i'w defnyddio. Mae llawer o gofion cyfieithu eraill, yn rhai perchnogol, a chod agored, yn bodoli. Mae'r Google Translator Toolkit [39] yn offeryn Cof Cyfieithu sy'n gweithio drwy fod y cyfieithydd yn uwchlwytho testun gwreiddiol; wedyn bydd y Toolkit yn rhannu'r dogfennau yn segmentau, yn rhoi cynnig ar gyfieithu'r segmentau hynny, ac yna yn dangos gweddau gwahanol o'r deunydd. Ar ochr chwith y sgrin dangosir y ddogfen wreiddiol, ar yr ochr dde dangosir y ddogfen sydd wedi ei rhag-gyfieithu. Mae ffenestr yn cael ei dangos hefyd lle y gellir golygu'r segment a rag-gyfieithwyd, neu (os na rag-gyfieithwyd) ei gyfieithu. Ar adegau rheolaidd, bydd y corpws o gyfieithiadau a gynhyrchir yn dilyn ymyrraeth y cyfieithydd dynol yn cael ei gynaeafu a'i roi i mewn i'r peiriant cyfieithu, sef y tanwydd y tu ôl i Google Translate, a thrwy hynny creir cylch rhinweddol o ansawdd cyfieithu. Po fwyaf y corpws o gyfieithiadau enghreifftiol, uchaf y bydd y tebygolrwydd ystadegol y bydd y cyfieithu awtomatig a gynigir o ansawdd uwch, heb fod arno angen cymaint o ôl-olygu ac yn y blaen. Mae'r posibiliadau ar gyfer rhannu cyfieithiadau, mewn amser go iawn, rhwng unrhyw nifer o sefydliadau, heb yr angen am meddalwedd ar ochr cleientiaid i gael eu gosod, yn aruthrol.

Mae nifer o ganolfannau yn bodoli, ledled y byd, sy'n ymchwilio i'r posibilrwydd o ddefnyddio posibiliadau corpora a chyfieithu trwy gyfieithu awtomatig hybrid mewn amgylchedd mainc gwaith. Ychydig iawn, fodd bynnag, sydd wedi ystyried anghenion y cyfieithwyr dynol eu hunain sydd yn defnyddio'r dechnoleg bob dydd. Mae un astudiaeth o'r fath [40] wedi gwneud hynny. Gallai mainc o'r fath gael ei defnyddio gan yr holl gyfieithwyr mewn sefydliad penodol, neu rwydwaith ehangach o sefydliadau a byddai'n cynnwys cyfleusterau o'r fath, yn ogystal ag awtomeiddio cyfieithu, er enghraifft, offer awduro rhagfynegol, didoli segmentau yn nhrefn y wyddor, a diweddaru mewn amser go iawn tuag at y cam cwblhau. Gall y technolegau cysylltiedig wella ymhellach cynhyrchedd cyfieithwyr, ac felly gwella tirwedd ieithyddol Cymru. Rydym yn cymryd rhan mewn chwyldro technolegol tawel a fydd, o gael buddsoddiad angenrheidiol mewn cydrannau technoleg iaith addas, yn trawsnewid cynllunio ieithyddol statws ar draws y byd.

4.3 CORPORA

Mae Corpora Cymraeg ysgrifenedig o faint digonol (a allai, er enghraifft, gynnwys nifer fawr o fersiynau electronig o gyhoeddiadau print) yn rhagofyniad ar gyfer datblygiadau pellach mewn technoleg iaith. Mae technoleg iaith o'r fath yn sail ar gyfer llawer o raglenni Cymraeg eraill, er enghraifft, y dechnoleg cyfieithu awtomatig a lleferydd a drafodir yn y papur gwyn hwn. A hithau ar flaen y gad yn y maes hwn ers blynyddoedd lawer, mae Canolfan Bedwyr (ac eraill) wedi datblygu lemateiddwyr (sy'n torri geiriau i lawr ac yn tagio eu ffurfiau gramadegol), corpora (cronfeydd data mawr o destun neu leferydd Cymraeg), algorithmau ar gyfer trefnau didoli a materion peirianneg iaith eraill. Er y bydd y rhan fwyaf o'r adnoddau hyn *ynddynt eu hunain* ond o ddiddordeb i arbenigwyr, mae effaith yr offer angenrheidiol hyn yn bellgyrhaeddol iawn ac yn arwyddocaol, wrth iddynt fwydo i mewn i raglenni technoleg iaith eraill. Gweler hefyd y dechnoleg awtoglosio y cyfeiriwyd ati uchod. Isod ceir rhestr (heb fod yn gynhwysfawr) o rai o'r corpora mwyaf arwyddocaol sydd ar gael ar gyfer y Gymraeg.

- *CEG* (Corpws Electroneg o'r Gymraeg)
- Sefydliad Cenedlaethol er Ymchwil i Addysg, *Ein Geiriau Ni*
- *Corpws Siarad* (Corpws Lleferydd).

4.4 ARGAELEDD OFFER AC ADNODDAU

Mae Ffigur 6 yn rhoi sgôr ar gyfer y gefnogaeth technoleg iaith sydd ar gyfer y Gymraeg. Mae'r graddio hwn o offer ac adnoddau cyfredol wedi'i gynhyrchu ar sail amcangyfrifon yn seiliedig ar raddfa o 0 (isel iawn) i 6 (uchel iawn) gan ddefnyddio saith maen prawf. Gellir crynhoi'r canlyniadau allweddol fel a ganlyn: Er bod sefyllfa'r Gymraeg yn *weddol* dda ymhlith ieithoedd lleiafrifol eraill o ran yr offer, a'r adnoddau mwyaf sylfaenol sydd ar gael, megis corpora, geiriaduron ffurfdroadol, toceneiddwyr, rhyngwynebau, tagwyr a lemateiddwyr, nid yw hyn yn rheswm i orffwys ar ein rhwyfau. Mae llawer o'r adnoddau presennol heb eu safoni felly mae angen mentrau i safoni'r data a fformatau cyfnewid, er enghraifft ym maes rheoli terminoleg a rhannu cofion cyfieithu. Hefyd ceir cynnyrch unigol a chanddynt swyddogaethau cyfyngedig mewn isfeysydd megis synthesis lleferydd, adnabod llais ac echdynnu gwybodaeth. Ar hyn o bryd, dim ond nifer fach o gwmnïau neu sefydliadau yng Nghymru sy'n gweithio ym maes technoleg iaith. Mae'n eithriadol o bwysig felly i barhau â chefnogaeth gyhoeddus i dechnoleg iaith ar gyfer y Gymraeg, yn arbennig o ystyried ehangu tirwedd ieithyddol y Gymraeg yn dilyn gweithredu Mesur y Gymraeg (Cymru) 2011. Mae'n arbennig o braf, felly, nodi'r alwad ddiweddar ar gyfer rhaglenni grant a estynnwyd gan Lywodraeth Cymru yn seiliedig ar ei Dogfen Strategaeth Technoleg Cymraeg a'r Cynllun Gweithredu cysylltiedig.

4.5 CYMHARIAETH DRAWSIEITHYDDOL

Mae cyflwr presennol y gefnogaeth technoleg iaith yn amrywio'n sylweddol o'r naill gymuned ieithyddol i'r llall. Er mwyn cymharu'r sefyllfa rhwng ieithoedd, bydd yr adran hon yn cyflwyno gwerthusiad yn seiliedig ar ddau faes rhaglen enghreifftiol (cyfieithu awtomatig a phrosesu lleferydd) ac un dechnoleg sylfaenol (dadansoddi testun), yn ogystal ag adnoddau sylfaenol sydd eu hangen ar gyfer adeiladu rhaglenni technoleg iaith. Mae'r ieithoedd wedi eu categoreiddio gan ddefnyddio'r raddfa pum pwynt canlynol:

1. Cefnogaeth ragorol
2. Cefnogaeth dda
3. Cefnogaeth gymedrol
4. Cefnogaeth ddarniog
5. Cefnogaeth wan neu ddim cefnogaeth

Cafodd cefnogaeth TI ei mesur yn ôl y meini prawf canlynol:

Prosesu Lleferydd: Ansawdd offer adnabod lleferydd cyfredol, ansawdd y technolegau synthesis lleferydd cyfredol, cwmpas y parthau, nifer a maint y corpora lleferydd cyfredol, maint ac amrywiaeth y rhaglenni lleferydd sydd ar gael.

Cyfieithu awtomatig: Ansawdd technolegau cyfredol cyfieithu awtomatig, nifer y parau iaith a gwmpesir, cwmpas y ffenomenau ieithyddol a meysydd, ansawdd a maint presennol corpora cyfochrog, nifer ac amrywiaeth y rhaglenni cyfieithu awtomatig sydd ar gael.

Dadansoddi Testun: Ansawdd a chwmpas technolegau dadansoddi testun cyfredol (morffoleg, cystrawen, semanteg), cwmpas y ffenomenau ieithyddol a meysydd, maint ac amrywiaeth y rhaglenni sydd ar gael, ansawdd a maint y corpora testun cyfredol (wedi eu hanodi), ansawdd a chwmpas gramadegau ac adnoddau geirfaol cyfredol.

Adnoddau: Ansawdd a maint y corpora testun cyfredol, corpora lleferydd a chorpora cyfochrog, ansawdd a chwmpas yr adnoddau a gramadegau geirfaol presennol. Dengys Ffigurau 7 i 10 fod Cymru yn y clwstwr is ar gyfer bron pob un o'r offer a'r adnoddau a restrir. Mae'n cymharu'n dda ag ieithoedd eraill sydd â nifer fach o

	Maint	Argaeledd	Ansawdd	Cwmpas	Aeddfedrwydd	Cynaliadwyedd	Modd Addasu
Technoleg Iaith: Offer, Technolegau a Rhaglenni							
Adnabod Lleferydd	1	1	1	1	1	1	3
Synthesis Lleferydd	1	2	2	2	3	2	3
Dadansoddi Gramadegol	2	1	2	2	3	2	1
Dadansoddi Semantaidd	2	2	2	2	2	2	2
Creu testun	2	2	2	2	2	2	2
Cyfieithu awtomatig	3	3	3	2	1	1	2
Adnoddau Ieithyddol: Adnoddau, Cronfeydd Data a Gwybodaeth							
Corpora testun	1	1	2	1	2	2	1
Corpora lleferydd	4	3	4	4	4	4	3
Corpora cyfochrog	3	3	2	3	3	4	3
Adnoddau geiriadurol	3	2	3	2	2	4	4
Gramadegau	4	3	3	3	3	5	4

6: Cyflwr cefnogaeth technoleg iaith i'r Gymraeg

siaradwyr, megis Estonia, Latfia, Lithwania, Slofacia. Fodd bynnag, mae'r holl ieithoedd hyn yn llusgo ymhell y tu ôl ieithoedd mawr fel Almaeneg a Ffrangeg, er enghraifft. Ond mae'n eglur nad yw hyd yn oed adnoddau ac offer TI ar gyfer yr ieithoedd hynny eto yn cyrraedd y safon a chwmpas adnoddau offer tebyg ar gyfer y Saesneg, sydd yn arwain bron ym mhob maes TI. Ac mae dal i fod digon o fylchau mewn adnoddau iaith Saesneg o ran rhaglenni o safon uchel.

4.6 CASGLIADAU

Mae'n amlwg y gall rhwystrau sosioseicolegol hanesyddol a diwylliannol rhag defnyddio'r iaith ddiglosig 'L' (y Gymraeg yn yr achos hwn) dal i fodoli wrth gyflwyno'r iaith i feysydd lle nad oedd yn bodoli o'r blaen—ac nid oedd disgwyl cyffredinol iddi fodoli. Ym mhob maes, mae newid arferion defnyddio'r iaith wedi bod yn rhan o bob dogfen strategaeth iaith ar ryw ffurf neu'i gilydd—mae'n weithgaredd sy'n cymryd amser hir ac mae angen i gynllunwyr iaith chwarae gêm hir. Yr hyn sy'n eglur, o'r gweithgarwch sylweddol ym myd Technoleg Iaith cod agored a pherchnogol, yw bod unigolion yn barod i ddefnyddio meddalwedd Cymraeg os yw ar gael yn hawdd, o ansawdd uchel, pan fydd lefel benodol o ymwybyddiaeth ohoni a phan fydd peth esbonio wedi bod arni. Cysyniad 'Cynnig Rhagweithiol', a ddaeth yn boblogaidd yn y lle cyntaf yng Nghanada, ac a ddaethpwyd i Ewrop gan brosiect *From Act to Action* [41] sydd ar ei fwyaf amlwg yn y maes hwn; fel yn achos yr holl wasanaethau eraill sy'n gysylltiedig ag iaith (os nad yw defnyddiwr yn ymwybodol ohoni, heb dderbyn cyn-

nig rhagweithiol o wasanaeth ieithyddol, sut (ac yn wir pam) y byddai'r person lleyg yn mynd allan o'i ffordd i ddod o hyd iddi a'i defnyddio?) Mae'r potensial ar gyfer normaleiddio'r Gymraeg ym maes technoleg iaith ac ar gyfer technoleg iaith i normaleiddio'r Gymraeg drwy gydlynu rhannu data cyhoeddus mawr yn enfawr.

Yn y gyfres hon o bapurau gwyn, yr ydym wedi gwneud ymdrech bwysig drwy asesu'r gefnogaeth technoleg iaith ar gyfer 31 o ieithoedd Ewropeaidd, a thrwy ddarparu cymhariaeth lefel uchel ar draws yr ieithoedd hyn. Drwy nodi'r bylchau, yr anghenion a'r diffygion, mae cymuned technoleg iaith Ewrop a'i rhanddeiliaid cysylltiedig bellach ar dir i gynllunio rhaglen ymchwil a datblygu fawr a chanddi'r nod o alluogi cyfathrebu gwirioneddol amlieithog, â chymorth technoleg ar draws Ewrop. Mae canlyniadau'r gyfres hon o bapurau gwyn yn dangos bod gwahaniaeth dramatig o ran cefnogaeth technoleg iaith rhwng y gwahanol ieithoedd Ewropeaidd. Er bod meddalwedd ac adnoddau o ansawdd da ar gael ar gyfer rhai ieithoedd a meysydd rhaglen eraill fel arfer bydd gan rai eraill—fel arfer y rhai llai eu maint—fylchau sylweddol. Nid oes gan sawl iaith dechnolegau sylfaenol ar gyfer dadansoddi testun a'r adnoddau hanfodol. Mae gan eraill offer ac adnoddau sylfaenol ond, er enghraifft, mae gweithredu dulliau semantig yn dal i fod yn bell i ffwrdd. Felly, mae angen ymdrech ar raddfa fawr i gyrraedd y nod uchelgeisiol o ddarparu cefnogaeth technoleg iaith o ansawdd uchel ar gyfer pob iaith Ewropeaidd, er enghraifft trwy gyfieithu awtomatig o ansawdd uchel. Nod hirdymor META-NET yw galluogi creu technoleg iaith o ansawdd uchel ar gyfer pob iaith. Mae hyn yn gofyn i'r holl randdeiliaid—mewn gwleidyddiaeth, ymchwil, busnes a'r gymdeithas—i uno eu hymdrechion. Bydd y dechnoleg a grëir o ganlyniad yn helpu i chwalu'r rhwystrau cyfredol ac yn adeiladu pontydd rhwng ieithoedd Ewrop, gan fraenaru'r tir ar gyfer undod gwleidyddol ac economaidd trwy amrywiaeth ddiwylliannol.

Cefnogaeth Ragorol	Cefnogaeth Dda	Cefnogaeth Gymhedrol	Cefnogaeth Ddarniog	Cefnogaeth Wan/dim
	Saesneg	Tsieceg Iseldireg Ffinneg Ffrangeg Almaeneg Eidaleg Portiwgaleg Sbaeneg	Basgeg Bwlgareg Catalaneg Daneg Estoneg Galisieg Groeg Hwngareg Gwyddeleg Norwyeg Pwyleg Serbeg Slovaceg Slovene Swedeg	Croatieg **Cymraeg** Islandeg Latfieg Lithwanieg Malteg Rwmaneg

7: Prosesu lleferydd: cyflwr cefnogaeth technoleg iaith i 31 iaith Ewropeaidd

Cefnogaeth Ragorol	Cefnogaeth Dda	Cefnogaeth Gymhedrol	Cefnogaeth Ddarniog	Cefnogaeth Wan/dim
	Saesneg	Ffrangeg Sbaeneg	Catalaneg Iseldireg Almaeneg Hwngareg Eidaleg Pwyleg Rwmaneg	Basgeg Bwlgareg Croatieg Tsieceg Daneg Estonieg Ffinneg Galisieg Groeg Islandeg Gwyddeleg Latfieg Lithwanieg Malteg Norwyeg Portiwgaleg Serbieg Slovaceg Slofeneg Swedeg **Cymraeg**

8: Cyfieithu awtomatig: cyflwr y gefnogaeth technoleg iaith i 31 iaith Ewropeaidd

Cefnogaeth Ragorol	Cefnogaeth Dda	Cefnogaeth Gymhedrol	Cefnogaeth Ddarniog	Cefnogaeth Wan/dim
	Saesneg	Iseldireg Ffrangeg Almaeneg Eidaleg Sbaeneg	Basgeg Bwlgareg Catalaneg Tsieceg Daneg Ffinneg Galisieg Groeg Hwngareg Norwyeg Pwyleg Portiwgaleg Rwmaneg Slovaceg Slofeneg Swedeg	Croatieg **Cymraeg** Estonieg Islandeg Gwyddeleg Latfieg Lithwanieg Malteg Serbieg

9: Dadansoddi testun: cyflwr y gefnogaeth technoleg iaith i 31 iaith Ewropeaidd

Cefnogaeth Ragorol	Cefnogaeth Dda	Cefnogaeth Gymhedrol	Cefnogaeth Ddarniog	Cefnogaeth Wan/dim
	Saesneg	Tsieceg Iseldireg Ffrangeg Almaeneg Hwngareg Eidaleg Pwyleg Sbaeneg Swedeg	Basgeg Bwlgareg Catalaneg Croatieg Daneg Estonieg Ffinneg Galisieg Groeg Norwyeg Portiwgaleg Rwmaneg Serbieg Slovaceg Slofeneg	Islandeg Gwyddeleg Latfieg Lithwanieg Malteg **Cymraeg**

10: Adnoddau lleferydd a thestun: cyflwr y gefnogaeth i 31 iaith Ewropeaidd

YNGLŶN Â META-NET

Mae META-NET yn Rhwydwaith o Ragoriaeth a ariennir yn rhannol gan y Comisiwn Ewropeaidd. Ar hyn o bryd mae'r rhwydwaith yn cynnwys 54 canolfan ymchwil mewn 33 gwlad Ewropeaidd. Mae META-NET yn sefydlu META, Cynghrair Technoleg Amlieithog Ewrop, cymuned gynyddol o weithwyr proffesiynol a sefydliadau technoleg iaith yn Ewrop. Mae META-NET yn meithrin y seiliau technolegol ar gyfer cymdeithas gwybodaeth Ewropeaidd wirioneddol amlieithog sy'n:

- gwneud cyfathrebu a chydweithio yn bosibl ar draws ieithoedd;
- caniat áu mynediad cyfartal i bawb yn Ewrop i wybodaeth waeth beth yw eu hiaith;
- adeiladu ar swyddogaethau technoleg gwybodaeth wedi'i rhwydweithio ac yn adeiladu arni.

Mae'r rhwydwaith yn cefnogi Ewrop sy'n uno ar ffurf marchnad ddigidol sengl ac fel gofod gwybodaeth. Mae'n ysgogi ac yn hyrwyddo technolegau amlieithog ar gyfer pob iaith Ewropeaidd. Mae'r technolegau hyn yn cefnogi cyfieithu awtomatig, creu cynnwys, prosesu gwybodaeth a rheoli gwybodaeth ar gyfer amrywiaeth eang o feysydd pwnc a rhaglenni. Maent hefyd yn galluogi rhyngwynebau greddfol sy'n seiliedig ar dechnoleg iaith yn amrywio o offer electroneg y cartref, i beiriannau a cherbydau i gyfrifiaduron a robotiaid. Lansiwyd META-NET ar 1 Chwefror 2010; ac mae eisoes wedi cynnal gweithgareddau amrywiol yn ei thair llinell gweithredu META-VISION, META-SHARE a META-RESEARCH.

Mae **META-VISION** yn meithrin cymuned rhanddeiliaid ddeinamig a dylanwadol sy'n uno o gwmpas gwledigaeth ranedig ac agenda ymchwil strategol gyffredin (SRA). Prif ffocws y gweithgaredd hwn yw adeiladu cymuned technoleg iaith gydlynol yn Ewrop drwy ddod â chynrychiolwyr o grwpiau rhanddeiliaid amrywiol ac wedi eu ffragmenteiddio. Paratowyd y Papur Gwyn cyfredol ynghyd â chyfrolau ar gyfer 30 o ieithoedd eraill. Datblygwyd y weledigaeth am dechnoleg wedi'i rhannu Grŵp Gweledigaeth sectoraidd. Sefydlwyd Cyngor Technoleg META er mwyn trafod a pharatoi'r SRA yn seiliedig ar y weledigaeth, a hynny wrth ryngweithio'n agos gyda'r gymuned technoleg iaith yn ei chyfanrwydd.

Mae **META-SHARE** yn creu cyfleuster agored, wedi'i ddosbarthu, ar gyfer cyfnewid a rhannu adnoddau. Bydd y rhwydwaith cymar-i-gymar o storfeydd yn cynnwys data, offer a gwasanaethau ieithyddol a gwasanaethau gwe wedi eu dogfennu gyda metaddata o ansawdd uchel ac wedi'i drefnu'n gategorïau safonol. Bydd modd hawdd ac unffurf o chwilio'r adnoddau hyn ac o gael mynediad iddynt. Mae'r adnoddau yn cynnwys deunyddiau rhad ac am ddim, cod agored yn ogystal ag eitemau cyfyngedig, sydd ar gael yn fasnachol, yn seiliedig ar ffioedd.

Mae **META-RESEARCH** yn codi pontydd i feysydd technoleg cysylltiedig. Mae'r gweithgaredd hwn yn ceisio godro datblygiadau mewn meysydd eraill gan fanteisio ar ymchwil arloesol a all fod o fudd technoleg iaith. Yn benodol, mae'r llinell gweithredu hon yn canolbwyntio ar gynnal ymchwil flaengar mewn cyfieithu awtomatig, casglu data, paratoi setiau data a threfnu adnoddau iaith at ddibenion gwerthuso; llunio rhestri o offer a dulliau, a threfnu gweithdai a digwyddiadau hyfforddi ar gyfer aelodau o'r gymuned.

office@meta-net.eu – http://www.meta-net.eu

1

EXECUTIVE SUMMARY

Language is the primary means of communication between humans. It allows us to express ideas and feelings, helps us to learn and teach, is essential for living, is the primary vehicle of transmission of culture, and is a symbol of identity. In our current level of globalization, we have many ways to easily communicate with people from all over the world. For example, the new information and communication technologies have enabled the development of social networks that have encouraged and enhanced interaction between people from virtually all countries and cultures. Also, in recent years, we have seen large movements of foreign people between our countries, i.e. tourism or immigration that creates the necessity for communication among different languages. This cross-lingual communication problem is often solved through the use of a lingua franca. The countries of Europe provide a clear example of linguistic and cultural diversity despite the fact that, during the last 60 years, Europe has increasingly become a distinct political and economic entity. As a result, language challenges are inevitably confronted by people in everyday life as well as in the spheres of business, politics and sciences.

Language challenges are inevitably confronted by people in everyday life as well as in the spheres of business, politics and sciences

The European Union's institutions spend about a billion Euros a year on maintaining their policy of multilingualism, i.e. translating texts and interpreting spoken communication. In parallel, English is becoming a lingua franca in the communication between European institutions and citizens. In the UK, as a case in point, we find a similar scenario. As so many public services are now provided either directly or indirectly by technological means, providing and recording Welsh language choice, and the language technology required for activating this choice is now a pressing issue for many reasons. The most salient justifications relate to the promotion of: active citizenship, equity of access to medical services, accessibility for those, for example, with impaired vision, and democratic representation itself.

Language technology may serve as a unique bridge between different languages

When combined with intelligent devices and applications, language technology will in the future be able to help citizens talk easily to each other and do business with each other even if they do not speak a common language. This, in the context of recently passed language legislation affecting public services in Wales is of paramount significance. Language technology solutions will eventually serve as a unique bridge between different languages. However, the language technologies and speech processing tools currently available on the market (ranging from question answering systems to natural language interfaces, and including translation systems and summarization tools, among many others), still fall short of this ambitious goal. As early as the late 1970s, the European Union realised the pro-

found relevance of language technology as a driver of European unity, and began funding its first research projects within this emerging field. At the same time, national and autonomic projects were set up that generated valuable results but never led to concerted European action. The predominant language technologies today rely on imprecise statistical approaches that do not make use of deeper linguistic methods, rules and knowledge. For example, sentences are automatically translated by comparing a new sentence against thousands of sentences previously translated by humans. The quality of the output largely depends on the size and quality of the available sample corpus. However, even this quasi-imprecise statistical method is far more productive than the labours of a lone translator not benefitting from such technology, or benefitting from real-time sharing of other translators' work via Translation Memory. While the automatic translation of simple sentences, in languages with sufficient amounts of available text material, can achieve useful results, such shallow statistical methods are doomed to fail in the case of languages with a much smaller body of sample material or in the case of sentences with complex structures. Analysing the deeper structural properties of languages is the only way forward if we wish to build applications that perform well across a wide range of languages. The solution to the cross-language communication problem is therefore to build key enabling technologies. To achieve this goal and preserve Europe's cultural and linguistic diversity, it is necessary to first carry out a systematic analysis of the linguistic particularities of all European languages, and the current state of language technology to support them. This is the purpose of the White Paper on Welsh.

The solution to the cross-language communication problem is to build key enabling technologies

As this series of white papers demonstrates, there is a dramatic difference between Europe's member states in terms of both the maturity of the research and in the preparedness with regard to recognising and implementing language solutions. One of the propositions and conclusions based on evidence of the offerings available is that Welsh is one of the EU languages that still needs further research before truly effective language technology solutions are ready for widespread everyday use, and that the language is normalised in technology, and that technology normalises the language to its full potential. Whilst language technology is an enabling technology and not an end in itself for the person-in-the-street, it is imperative that smaller languages such as Welsh receive due attention or their speakers will be further disenfranchised.

It is imperative that minority languages receive due attention, or their speakers will be further disenfranchised

Language technology also has a great role to play in terms of recording citizens' language choice. This ideal enabling situation would be for the technology to enable those agents of the state and other sectors to *proactively offer* services in Welsh, because citizens' language choice will already be known to them. Whilst the legislative situation in Wales is developing its discourse for the state to be a provider of such services, language technology will enable equity of language provision for all citizens (at its most simple level routing Welsh-speaking phone calls automatically to Welsh speaking staff, matching Welsh speaking social services/medical staff to Welsh speaking service users and so on), at its most complex, automatically translating documents and meetings.

Technology will provide fairness in terms of language provision for all citizens

2

LANGUAGES AT RISK: A CHALLENGE FOR LANGUAGE TECHNOLOGY

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in digital information and communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

The digital revolution is comparable to Gutenberg's invention of the printing press

After Gutenberg's invention of the press, real breakthroughs in communication and knowledge exchange were accomplished by efforts such as the translation of the Bible into vernacular languages. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries
- the teaching and translation of languages enabled exchanges across languages; the creation of editorial and bibliographic guidelines assured the quality of printed material
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

Likewise, in the past twenty years, information technology has helped to further automate and facilitate language processing and knowledge exchange:

- desktop publishing software has replaced typewriting and typesetting;
- Microsoft PowerPoint, and latterly, Prezzi have replaced overhead projector transparencies;
- e-mail allows documents to be sent and received more quickly than using a fax machine;
- Skype and other offerings provide cheap internet phone calls and host virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- web search engines provide keyword based access;
- online services like Google Translate produce quick though approximate translations;
- social media platforms such as Facebook and Twitter facilitate communication, collaboration, and information sharing.

Although these tools and applications are helpful, they are not yet capable of supporting a fully sustainable, multilingual European society in which information and goods can flow freely.

2.1 LANGUAGE BORDERS HOLD BACK THE EUROPEAN INFORMATION SOCIETY

We cannot predict exactly what the future information society will look like. But there is a strong likelihood that the revolution in communication technology is bringing people speaking different languages together in new ways. This is putting pressure on individuals to learn new languages and especially on developers to create new technology applications to ensure mutual understanding among speakers of different languages and access to shareable knowledge. In a global economic and information space, more languages, speakers and content interact more quickly with new types of media. The current popularity of 'Web2.0' social media (such as Wikipedia, Facebook, Twitter, YouTube, etc) is only the tip of the iceberg. Content is king and the user is now in control to an extent never before witnessed. This freeing of control is akin to the revolution in print capitalism which helped the formation of the monolingual, monolithic nation state three centuries ago. However, this time *individuals* are in control, and their languages have an avenue for expression hitherto denied to them. Could Web2.0, connected with relevant language technology, be the new print capitalism, historically used as a homogenising force by nation states, but this time *favouring* RMLs?

In a worldwide economic and information space, more languages, speakers and content are interacting more quickly with more types of media

Today, we can transmit gigabytes of text around the world in a few seconds before we recognise that it is in a language we do not understand. According to a report from the European Commission [2], 57% of internet users in Europe purchase goods and services in non-native languages (English is the most common foreign language followed by French, German and Spanish). 55% of users read content in a foreign language while only 35% use another language to write e-mails or post comments on the web. A few years ago, English might have been the lingua franca of the web. The vast majority of content on the web was in English, possibly because of its initial development in English-speaking countries. In terms of the early-stage development of web-content for RML languages, one postulate is that this may have been because of the diglossia-provoked mentalities and attitudes described above.

The omnipresent digital divide caused by language borders has not received much public attention

Fortunately, this situation has now drastically changed. The amount of online content in other European (as well as Asian and Middle Eastern) languages has exploded. Surprisingly, this ubiquitous digital divide due to language borders has not gained much public attention; yet, it raises a very pressing question: How *exactly* can technologies enable languages to thrive in the networked information and knowledge society when so many of them, the world over, are under threat? For example, Ethnologue [3] notes that there are approximately 7,015 languages in the world. The majority of these are under threat and will cease to be transmitted to future generations (such intergenerational language transmission being one of the key indicators of language vitality). Such risks are the focus of the next section of this white paper.

2.2 OUR LANGUAGES AT RISK

While the printing press and associated print capitalism helped step up the exchange of information in Europe [4], it contributed substantially to the process by

which many European languages lost substantial numbers of speakers. Many regional or minority languages were rarely printed and were limited to oral forms of transmission, which in turn restricted their scope of use. Welsh benefited from the standard written language created by Bishop William Morgan's 1588 translation of the Bible. But how will Welsh survive the impact of the internet? Europe's approximately 80 languages are one of its richest and most important cultural assets [5], and a vital part of its unique social model. While languages such as English and Spanish are likely to survive in the emerging digital marketplace, many European languages could become irrelevant in a networked society unless sufficient strategic steps are taken. This would weaken Europe's global standing, and run counter to the strategic goal of ensuring equal participation for every European citizen regardless of language. The wide variety of languages in Europe is one of its richest and most important cultural assets. According to a UNESCO report on multilingualism [6], languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society.

How will the Welsh language survive the effect of the Internet?

We now move from the political, philosophical, historic and strategic, to the operational. To a certain degree, and to grossly oversimplify by neglecting the social context outlined above, language (reified as a manipulable construct in itself outside of that social context for the purpose of ICT) *could*, simplistically, be portrayed as nothing more than 'content.' In summary, technology and language technology have a large role to play in helping people's bilingual lives and to be used as an instrument of status planning to change deeply-embedded attitudes toward the use of the 'L' language in hitherto unfamiliar domains. Technology is omnipresent. Multilingualism must be omnipresent in technology.

2.3 LANGUAGE TECHNOLOGY IS A KEY ENABLING TECHNOLOGY

In the past, investment efforts in language preservation focused on language education and translation. According to one estimate [7], the European market for translation, interpretation, software localisation and website globalisation was €8.4 billion in 2008 and was expected to grow by 10% per annum. Yet this figure covers just a small proportion of current and future needs in communicating between languages. The most compelling solution for ensuring the breadth and depth of language usage in Europe tomorrow is to use appropriate technology, just as we use technology to solve our transport, energy and disability needs among others. Language technology targeting all forms of written text and spoken discourse can help people to collaborate, conduct business, share knowledge and participate in social and political debate regardless of language barriers and computer skills.

Technology is omnipresent. Multilingualism must be omnipresent in technology

It often operates invisibly inside complex software systems to help us already today to: find information with a search engine; check spelling and grammar in a word processor; view product recommendations in an online shop; follow the spoken directions of a navigation system; translate web pages via an online service. Language technology consists of a number of core applications that enable processes within a larger application framework. The purpose of the META-NET language white

papers is to focus on how ready these core enabling technologies are for each European language. Europe needs robust and affordable language technology for all European languages.

Europe needs robust, affordable language technology for all its languages

To maintain our position in the frontline of global innovation, Europe will need language technology, tailored to all European languages, that is robust and affordable and can be tightly integrated within key software environments. Without language technology, we will not be able to achieve a really effective interactive, multimedia and multilingual user experience in the near future.

2.4 OPPORTUNITIES FOR LANGUAGE TECHNOLOGY

In the world of print, the technology breakthrough was the rapid duplication of an image of a text using a suitably powered printing press. Human beings had to do the hard work of looking up, assessing, translating, and summarising information. Language technology can now simplify and automate the workflow processes of translation, content production, and knowledge management. It can also empower intuitive speech-based interfaces for household electronics, machinery, vehicles, computers and robots. Real world commercial and industrial applications are still in the early stages of development, yet R and D achievements are creating a genuine window of opportunity. For example, machine translation (described below in the case of Welsh) is already reasonably accurate in specific domains, and experimental applications provide multilingual information and knowledge management, as well as content production, in many European languages. As with most technologies, the first language applications such as voice based user interfaces and dialogue systems were developed for specialised domains (for example, specific narrow fields of medicine, for note taking), and often exhibit limited performance. However, there are huge market opportunities in the education and entertainment industries for integrating language technologies into games, edutainment packages, libraries, simulation environments and training programmes. Mobile information services, computer assisted language learning software, eLearning and blended learning environments, self-assessment tools and plagiarism detection software (used to detect plagiarised work on submission of students' assignments) are just some of the application areas in which language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse. The same goes for 'big data' monitoring [8] and trending research projects, such as those used by the security services to detect possible social unrest and 'hate speech'. Language technology helps overcome the 'disability' (as perceived by some) of linguistic diversity. Language technology represents a tremendous opportunity for the European Union. It can help to address the complex issue of multilingualism in Europe 'the fact that different languages coexist naturally in European businesses, organisations and schools. However, citizens need to communicate across the language borders of the European Common Market, and language technology can help overcome this final barrier, while supporting the free and open use of individual languages. Looking even further ahead, innovative European multilingual language technology will provide a benchmark for our global partners when they begin to support their own multilingual communities.

Language technology helps overcome the 'disability' (as some would call it) of linguistic diversity

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. Widely used technologies, such as the spelling and grammar correctors in word processors, are typically monolingual, and are only available for a handful of languages. Online machine translation services, although useful for quickly generating a reasonable approximation of a document's contents, are fraught with difficulties when highly accurate and complete translations are required. The current pace of progress in language technology is too slow. Due to the complexity of human language, providing for the computational modelling of our tongues and testing it in the real world is a long, costly business that requires sustained funding commitments. Europe must therefore maintain its pioneering role in facing the technological challenges of a multiple language community by inventing new methods to accelerate development right across the map.

2.5 LANGUAGE ACQUISITION IN HUMANS AND MACHINES

To illustrate how computers handle language and why it is difficult to program them to process different tongues, let us look briefly at the way humans acquire first and second languages, and then see how language technology systems work. Babies acquire a language by linguistic interaction and by listening to the real interactions between their parents, siblings and other family members. From the age of about two, children produce their first words and short phrases. This is only possible because humans have a genetic disposition to imitate and then rationalise what they hear. Learning a second language at an older age requires more cognitive effort, largely because the person is not immersed in a language community of native speakers. At school, foreign languages are usually acquired by learning grammatical structure, vocabulary and spelling using drills that describe linguistic knowledge in terms of abstract rules, tables and examples. Humans acquire language skills in two different ways: learning from examples and learning the underlying language rules. Moving now to language technology, the two main types of systems 'acquire' language capabilities in a similar manner. Statistical (or data driven) approaches obtain linguistic knowledge from vast collections of concrete example texts or 'corpora'. While it is sufficient to use text in a single language for training, say, a spell checker, parallel texts in two or more languages have to be available for training a machine translation system. The machine learning algorithm then 'learns' patterns in terms of how words, short phrases and complete sentences are translated. This statistical approach usually requires millions of sentences to boost performance quality. This is one reason why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, and services such as Google Search or Google Translate, all rely on statistical approaches. The great advantage of statistics is that the machine learns quickly in a continuous series of training cycles. Another approach to language technology and to machine translation in particular, is to build rules-based systems. Experts in the fields of linguistics, computational linguistics and computer science first have to encode grammatical analyses (grammar rules) and compile vocabulary lists (lexicons). This is very time consuming and labour intensive. Some of the leading rules-based machine translation systems have been under constant development for more than 20 years. The great advantage of rules-based systems is that the experts have more detailed control over the language processing.

This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rules-based systems are used for language learning. However, due to the high cost of this work, rules-based language technology has so far only been developed for a few major languages. As the strengths and weaknesses of statistical and rules-based systems tend to be complementary, current research focusses on hybrid approaches that combine the two methodologies. The possibilities of a *tripartite* machine translation engine for Welsh (rules-based, statistical-based and example-based), combined with translation memory and wide real-time sharing of translations are discussed below. It should be noted, however, that these hybrid approaches have so far been less successful in industrial applications than in the research lab. The two main types of language technology systems acquire language in a similar manner as humans do. As we have seen in this chapter, many applications widely used in today's information society rely heavily on language technology, particularly in Europe's economic and information space. Although this technology has made considerable progress in the last few years, there is still huge potential to improve the quality of language technology systems. In the next chapters, we describe the role of the Welsh language in European information society and in the world, and assess its sociolinguistic background and the current state of language technology for Welsh.

3

THE WELSH LANGUAGE IN THE EUROPEAN INFORMATION SOCIETY

3.1 GENERAL FACTS

The Welsh language, a Celtic language (related to Breton, Cornish, Irish, Scottish Gaelic and Manx) has a similar post-enlightenment history to many other of the other regional or minority languages all over Europe, i.e. centrist policies and a sharp demographic decline beginning at the start of the twentieth century, followed by a slowing in that decline in the 1970s. In the case of Welsh, the latest decennial census results (2011) [9] found that 19% (562,000) of the population of Wales noted an ability to speak Welsh, a decrease in absolute numbers of 20,400 on the previous census (2001). It should be noted however, that the 2001 census had itself shown an increase of 80,000 people on the 1991 census. It would therefore appear, from the large-scale sample that the census provides, that the Welsh language is in a far safer situation in terms of its top level demographic figures than many of the other languages in the world. But such top level figures can, of course, only be superficial. The picture itself is much more complex [10].

This complexity manifests itself in the evolving age profile of Welsh speakers. For example, of the 10-14 year old age group, 42.2% were recorded as Welsh speakers at the 2011 census, higher than the figures for a century earlier in 1911 (and much higher than the 16.2% of those older than 65 years of age who noted that they spoke Welsh in 2011). Indeed, the most recent analysis of the 2011 Census figures for the Welsh language [11] shows that 'the number of children speaking Welsh is more than twice that of those aged 16-64 and the over 65s'. The Welsh Language Board also published a 'reasonable estimate' [12] that there could be 110,000 Welsh speakers living in England. The positive implications of language technology and technology in general for language use and maintenance in such diaspora communities are substantial. The Welsh language television channel, S4/C, broadcasts in England (and the rest of the UK outside Wales in Welsh). Of all those Welsh speakers (100%) 44.9% live in homes where everyone speaks Welsh (2001 Census figures) the corresponding analysis for 2011 figures is not available at the time of writing this white paper). These statistics mean that the family lives of very many Welsh speakers are bilingual, which pose particular challenges in terms of switchable interface language technology, for language status planning (i.e. how can people with differing language abilities in the same family or work unit share a computer if its interface is in the 'L' (diglossic lower status) language. It is therefore of fundamental importance for applied language planning that as much as possible of the technology people use in school and beyond is freely available in Welsh, and implemented in the organisations and networks that are used by particular target segments. It is beyond the scope of this white paper to provide a detailed statistical breakdown by age of Welsh language use and ability. However, it is worth noting that the more fluent a speaker is in Welsh, the more likely that speaker is to use the language every day.

There are substantial positive prospects for language technology and technology in general for language use and for sustaining fragmented linguistic communities

The former Welsh Language Board's Welsh Language Use Surveys (2004-2006) (the most recent national language use figures available) noted that of the 588,000 people it estimated were speakers of Welsh, 58% (317,000) considered themselves fluent. 76% of fluent speakers said they spoke Welsh every day, and Welsh was the language of the most recent conversation of 59% of the fluent speakers. During the life of the statutory Welsh Language Board (1993-2012), the discourse in Language Planning evolved from an overly simple desire merely to increase numbers able to speak Welsh to a more sophisticated behaviour-changing strategy using lessons learnt from the National Health Service's Public Health Promotion work (cf the Twf ['growth'] project to persuade parents to speak Welsh to their children where they may not have previously had enough confidence to do so) and to increase use of the Welsh language in diglossic 'H' situations via a project called *Mae gen ti ddewis...* ('You've got the choice'). The discourse of *use* rather than increasing numbers also prevails in more recent Welsh Government strategy documents.

Language use patterns are deeply rooted in social psychology, self-perception and perceptions of linguistic self-efficacy (whatever the true fluency of a given Welsh speaker)

Language behaviour patterns are deeply rooted in social psychology, self-perception, self-confidence and perceived linguistic self-efficacy (regardless of a Welsh speaker's actual language fluency) [13]. These elements come up time and time again in the research commissioned by the former Welsh Language Board. In short, perception is one's own subjective reality and, in combination with other factors, one acts within the parameters that one's self-belief creates. This is particularly salient in considering the use of language technology.

3.2 PARTICULARITIES OF THE WELSH LANGUAGE

Welsh has intrinsic linguistic features which make it dissimilar to many of the other languages of this white paper series (apart from its cousin language, Irish). This may make development, and in particular, cross-fertilization of language technology more challenging than, say, between French, Spanish and English.

There are 29 letters in the Welsh alphabet (Roman script is used). Welsh does not use 'x' or 'z', and 'j' is only normally used in words borrowed from English. The language benefits from full Unicode support and therefore, as a general rule, there should be no problem in depicting the characters used in any Unicode-compliant setting. One of the main particularities of Welsh is that is uses digraph letters (two symbols to produce a specific sound). These are ch, dd, ff, ng, ll, mh, nh, ph, rh, th. This poses a challenge to those technologists implementing, for example, sorting in databases, as 'Llandeilo' will come after 'Luton' (both place names).

Welsh, like its Irish cousin, is an inflectional language which means that its linguistic forms change depending on (for example) tense, number, and person. Take, for example the regular verb 'canu' (to sing). In compact, literary form, the verb could be conjugated as follows:

- Canaf
- Ceni
- Cân
- Canwn
- Canwch
- Canant

However, the gulf between formal written and spoken Welsh is wide. The oral language tends towards use of a more periphrastic structure (which itself is not unproblematic due to substantial dialectal variation in that periphrastic structure). By way of example, 'I am singing/I sing' in the periphrastic form could be heard as follows (depending on the form the speaker adopts):

- Dwi'n canu
- Rwy'n canu
- Rwyf yn canu
- Rydw i'n canu
- Fi'n canu (stigmatised form, but becoming more common)

The same variation also exists for other persons. The lack of a standard oral language causes problems, for example, for development of speech recognition systems, or indeed automatic translation or phonic searching. The relevant lemmatizing databases will also have to deal with this problem.

Another feature of the Celtic languages is their system of initial consonant mutation. Nine letters of the Welsh alphabet mutate, and there are three types of mutation, as shown in the table below. However, language technology needs to take into account the rules by which these mutations occur. For example, the word 'a' can mean 'and', or can be a preverbal particle. In the first instance, it would cause an aspirate mutation, in the second a soft. The object of compact form verbs also takes a soft mutation in Welsh. So, for example, 'Gwelodd fachgen' (He/She saw a boy) is only different from 'Gwelodd bachgen' (A boy saw) by the mutation. Placenames take a nasal mutation after 'yn' ('in'), but 'yn' can also be a preverbal particle. How will language technology be able to differentiate between these situations? This, of course, is also salient for the machine translation discussed later in this white paper.

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is slow for many of Europe's smaller languages. Europe must therefore maintain its pioneering role in facing the technology challenges of multiple language communities by inventing new methods to accelerate development right across the map. These could include both computational advances and techniques such as crowdsourcing. In the particular case of Welsh, some of the challenges that that need to be overcome in the development of language technology are as follows: Dialectal variation in Welsh is great, although all dialects are mutually intelligible. Up until the advent of S4/C, the Welsh language television channel in 1982, which has facilitated internal comprehension in the Welsh language speech community, many anecdotal complaints were heard regarding 'incomprehensibility' of Welsh dialects. This poses obvious challenges to technology, such as speech recognition. Welsh has benefitted from a standard written language since the translation of the Bible by Bishop William Morgan in 1588. This did much to conserve language unity, where, for example, other languages, such as Breton, fragmented dialectally. Religion has played, up until very recently, a large sociological role in language conservation in Wales, when the state was not as amenable to additive bilingualism as it now is.

Components could be shared and reused between languages

However, the standard written language described above is widely different from oral Welsh (in all its dialectal forms). So, for example, in the case of speech to text recognition used in a formal meeting, what language technology tools would be needed to allow one to transcribe the utterances to create a formal record in written Welsh of that meeting (without large amounts

	Soft	Nasal	Aspirate
T	D	NH	TH
C	G	NGH	CH
P	B	MH	PH
B	F	M	-
D	DD	N	-
G	[Disappears]	NG	-
M	F	-	-
RH	R	-	-
LL	L	-	-

of post-editing)? To illustrate this point, Professor Bobi Jones, who has written extensively on Welsh language matters, wrote of formal Welsh [14]: 'We speak the language of our parents and write the language of the grandparents of our great great grandparents!' At present, there is scant formal provision for the training of Welsh speakers as computational linguists. The requisite skillsets in technology, programming and correct Welsh language skills and computational linguistics are rare. It is often difficult to appoint staff even to content *creation* posts (for example, for websites) with sufficient *copywriting* and language skills.

Cost savings, increases in consistency and a decrease of repetitive translation can all be achieved via the sharing of Translation Memories

Europe-wide projects such as META-NET show the large economies of scale that can be found by tackling similar problems of many languages in one space. This is undoubtedly an opportunity for Welsh, and other smaller languages. And whilst Europe's many languages are not grammatically similar, some components may be shared and redeveloped between those languages which *are* related. Many generic technologies also should be fully utilised and, possibly, their architecture shared for Europe's languages such as:

- Cloud-based translation memories
- Hybrid machine translation (patched into the translation memories described in detail in this white paper)

In terms of the opportunities for the Welsh language that language technology may provide, we may count:

- The social capital that may accrue from crowdsourcing projects in the voluntary sector, and the social capital from joint origination of content in Welsh
- Savings in cost, large increase in consistency and a reduction of repetitive translation via large-scale sharing of Translation Memories in the public sector
- Speech synthesis may provide assistance for those learning Welsh who do not have every day access to speakers of the language
- In terms of the inclusivity agenda, Welsh language screenreaders will increase access of those with visual impairments to Welsh language content and services.
- It may also assist to post edit machine translated output more quickly than via keyboard

3.3 RECENT DEVELOPMENTS

The philosophy of post-enlightenment romantic nationalism manifested itself in the creation of monolithic

nation states which, in order to be monolithic, implemented centrist, uniforming policies which decreased (or in some cases, deleted) internal linguistic diversity. The sociopsychological aspects of the resultant diglossia for technology and 'modern' functions have been a fetter on Welsh language normalisation, particularly in 'new' spheres such as language technology. However, from the last third of the 20th century, a world-wide movement for *glocalisation* developed, prizing the local yet with a global worldview. This saw resurgence in interest in regional or minority languages. In the case of Welsh, wide-spread civil disobedience and non-violent protest (mostly, in the earlier years, on behalf of the Welsh Language Society) elicited government concessions. Thereafter other legislation followed:

- The Welsh Language Act (1967), allowing ministers to prescribe official versions of forms in Welsh, limited use of Welsh in the courts system and several other provisions.
- The Broadcasting Acts (1981 and 1982) which led to the establishment of S4/C, the Welsh language television channel, (S4/C is now a pioneering digital multi-platform broadcaster involving audiences in its programmes via social media)
- The Education Act 1988 which made Welsh a core subject in the National Curriculum in Wales (the language was not hitherto compulsorily taught in the school system, and many schools did not teach it at all).
- The Advisory Welsh Language Board (1988) which was to make recommendations to Ministers regarding the appropriacy of legislation for the Welsh language. One such recommendation caused the drafting of the Welsh Language Act 1993. This Act, which repealed many of the provisions of the 1967 Act and ensured that public organisations in serving the public in Wales, wherever they were situated in the UK, provided service to the public on a basis of equality between Welsh and English. This was ensured by 'Welsh language schemes', documents tailored to each individual organisation's circumstances. The Act established the statutory Welsh Language Board to 'promote and facilitate the use of the Welsh language'. The Board was abolished in March 2012, by the legislation described below.

- The Government of Wales Acts (1998, and 2006), granting devolution of power in limited fields to the National Assembly for Wales, and giving the Assembly the ability to do 'anything within its power' to promote the Welsh language.

- Other significant developments include civil society movements such as Mudiadau Dathlu'r Gymraeg (an umbrella movement of Welsh language organisations), Dyfodol i'r Gymraeg, (a lobby of eminent Welsh speakers) and Yr Awr Gymraeg (the Welsh Language Hour), during which Welsh speakers are once per week encouraged to use Welsh on Twitter for a given hour each week.

- A policy observatory, announced in January 2013 by the Welsh Language Commissioner, to study the implications of the Welsh language in every policy sphere.

- The Coleg Cymraeg Cenedlaethol (National Welsh Language College), a Government-funded, university-level Virtual organisation, coordinating Welsh-medium provision in Wales' higher education establishments. Lectureships in technology have been advertised.

Many of the above developments, to a degree, use, or created a need for language technology.

3.4 LANGUAGE PROMOTION AND REGULATION

One of the most significant developments in language policy in Wales was been the granting of Legislative Competence to the National Assembly for Wales to enact legislation in matters relating to certain aspects of the Welsh language. This enabled the Welsh Government to bring forward one of the main pillars of its 'One Wales' Coalition Agreement, i.e. to draft the Welsh Language (Wales) Measure (2011) [42]. The Measure gave Welsh Ministers the power to abolish the Welsh Language Board and reorganise its functions. The post of Welsh Language Commissioner was one of the mainstays of this legislation. The Commissioner came into being in April 2012. The Commissioner has power to enforce legal compliance with a new series of 'Language Standards', which are intended to replace the 541 extant Welsh Language Schemes, the main instrument of the current Welsh Language Act. These standards are subdivided as follows:

- Service delivery standards
- Policy making standards
- Operational standards
- Promotion standards
- Record keeping standards

The first three are most relevant to language technology. They ensure that services provided to the public in Wales (by whatever means) are provided via the language of the end user's choice and that that choice must be captured and reused via the whole dealings of organisation 'X' with that end user (websites, app, CRM etc).

Parts of the telephony sector may be brought under the aegis of the Welsh Language Measure, which therefore may mean compulsion to create mobile telephony interfaces in Welsh

Policy making standards will ensure that in every policy decision an organisation makes, the Welsh language must be considered as a factor. E.g. in terms of an organisation's long-term IT strategy, 'is system X available, or likely to be available in Welsh, and does it capture and manage language preference' ' if not, disregard, or supply a case for continuing with procurement (keeping due records of the reasoning behind the decision to continue). The operational standards regard the internal workings of a given organisation, and the rights that may arise from these standards would enable workers to communicate with each other in Welsh without let or hindrance and fully supported by statutory underpinning. Technology is a key facilitator for this and active offer, for example, of service or provision (administrative and content side) will be expected in applications such as content management systems and IT interfaces. The Measure also allows for major elements of the telecommunications sector to be brought within its purview at a later date thereby requiring, for example, interfaces of mobile to be available in Welsh.

Lastly, as strategic background to the policy framework effecting language technology in Wales, the Welsh Government has published a substantial Language Strategy document [16] (having already published a Strategy document for Welsh Language Education) published under the title of *Iaith Fyw: Iaith Byw* (A Living Language, a Language for Living). The strategy outlines the Government's vision for the bilingual Wales it wishes to see in the future, and Ministers have stated several times in public that one of the main philosophical foundations of that strategy will be increasing use of language provision which is already available by whatever means. This philosophy is outlined, and examined, in the context of IT provision which is already available in Welsh, below. The strategy envisions a 'strong representation of the Welsh language throughout the digital media' and devotes one of its six chapters to 'Infrastructure' for the

language, under which language technology is tackled. These are that chapter's strategic targets, noting that the Welsh Government will act by:

- encouraging major private sector service providers, including banks, retailers, mobile phone companies, software and hardware developers, and others to develop online services, applications and interfaces through the medium of Welsh
- facilitating the development of Welsh interfaces for commonly used social networking media, including open source software
- providing, possibly on a matched basis, seedcorn funding for initiatives such as these on an incremental basis over time
- developing a consensus around priority areas where technological investment is required

The Ministerial Task and Finish Group on Welsh Language Technology and Digital Media was formed in early 2012 and met several times to discuss a draft Strategy and Action plan for the Welsh language and technology. The strategy [17] deals in detail with all the above themes, and provides a welcome emphasis on content creation, which will, of course, strongly dovetail with the language technology tools and themes described below in the case of Welsh. The strategy proposes five fields of action:

- Marketing and awareness raising
- Influencing large software companies
- Encouraging development of new software packages and Welsh medium digital services
- Encouraging creation and sharing, and use of Welsh language digital content
- Sharing best practice in the public, private and third sectors.

These will be operationalized by Governmental funding, encouragement and legislative initiatives. In order to realise the Government's vision of a bilingual Wales, Welsh must have its rightful place in the world of technology, and a strategic, long-term approach is in place to ensure this. Other considerations influence the role and legal position of the Welsh language. Within the past decade it has become part of the equalities agenda yet it does not figure prominently in comparison with other equality strands based upon race, gender, sexual orientation or disability.

One way of contributing to this omnipresence [of Welsh in technology] has been to publish detailed, technical guidelines which would provide advice on how to create multilingual software or websites, emphasising the need for easy language switching. In April 2006, the Welsh Language Board therefore launched comprehensive *Bilingual Software Guidelines and Standards* [18], (on the same day as the first Strategy document for IT and the Welsh Language) [19] and circulated them to all organisations with a statutory Language Scheme under the Welsh Language Act 1993. It then held several seminars for technical practitioners to publicise the standards, one each in north and south Wales, and another in London (principally for Crown Bodies with a Language Scheme under the Welsh Language Act 1993). It was hoped that the advice this document provided, and the monitoring frameworks used to track its implementation, would improve the bilingual provision of electronic services of all kinds, improving on the performance noted in the two *Snapshot Surveys* of public sector websites the Board carried out in 2001 [43] and 2003 [21]. In August 2009, the Board launched a technical Accreditation Scheme [22] for the Standards Document on its website. This scheme, aimed at technically skilled IT staff within organisations with a Language Scheme under the Welsh Language Act 1993 (and any other organisation wishing to provide IT services bilingually), turns the *Bilingual Software Guidelines and Standards* into question form, enabling those

technical staff to ascertain easily whether a given system is compliant and supplies a language choice on a basis of equality between Welsh and English. The intention is that the results of these questions will be used as diachronic indicators for improving bilingual IT services. This document was updated, and reworked as one of the first guidance documents published by the Welsh Language Commissioner in 2012. All the Commissioner's guidance documents will, in due course, be published as Codes of Practice under the Welsh Language (Wales) 2011 Measure, with organisations having to prove how they have given 'due regard' to these codes in complying with the standards regime described above.

3.5 LANGUAGE AND TECHNOLOGY IN EDUCATION

The field of end user training and capacity building through the medium of Welsh in the field of IT requires specific attention. Various training suites exist, both in electronic and paper format, amongst them *The European Computer Driving Licence* [23] and Microsoft's *Digital Literacies* project, using the standardised terminology and screengrabs of the Welsh Language versions of its own products. As the world of IT moves so quickly, updates to this type of training will regularly be needed, and other training packages are in preparation. In terms of training for organisations to provide language technology, The Welsh Language Board circulated an Advice Note under the Welsh Language Act 1993, to all 541 organisations with a language scheme, explaining the current provision in terms of language technology, noting the myths outlined in this white paper and debunking them. Undergraduate university-level courses are available in language technology to differing degrees in different institutions, but not in all.

3.6 INTERNATIONAL ASPECTS

The Welsh Language is covered by provisions in the Framework Convention for the Protection of National Minorities and the UK Government has ratified 52 Clauses of the European Charter for Regional or Minority Languages for Welsh. It is also represented on the Network for the Promotion of Linguistic Diversity, the head office of which is in Cardiff.

3.7 WELSH ON THE INTERNET

Welsh Government figures [24] for 2012 show 70% of Welsh households had access to the internet. This equates to approximately 77% of people aged 18 or over having access to the internet at home. 73% of people said that they used the internet at home, work or elsewhere; this varied by age with a far greater proportion of people under 45 using the internet than those aged 45 and over. internet users under 25 years old are more likely (41%) than internet users aged 65 and over (8%) to have accessed the internet at another person's home in the last three months. Later figures show that the percentage of the Welsh Population that had *never* used the internet was 17.5% (compared with 14% [25] of the UK population as a whole. However, no accurate figures are available as to which language Welsh people use to access the internet. Wikipedia has Welsh as its 65th strongest content language (with *circa* 49,000 articles) [26], proving the vitality and energy of the open source community of volunteer content creators. Several browsers already have an interface for Welsh (Internet Explorer, Opera, Firefox), and several more, although not having Welsh interfaces, allow the browser *locale* to be switched to Welsh in order to automate presentation of Welsh content. Unfortunately, in order to avail oneself of this provision, one has to know what *locale* is, to know where it is, to want, and know how to change it, and to change it back if, for example, a soft-

ware update resets it. All this is independent of the language interface of a given operating system, for example, Microsoft Windows. A Welsh language interface has been developed for Google (search) and Gmail. Web search is the most commonly used internet application and it is ubiquitous across all sorts of devices and platforms. In addition, web search itself employs, or can employ, a range of language technologies (of various levels of sophistication) to improve results and overall quality. Aside from the prestige of being associated with such a successful international brand, Google's Welsh language interfaces provides not only a language appropriate internet experience for Welsh speakers, but it also reflects the growing need for appropriate search and language processing tools and services to deal with Welsh language data. The next section gives an introduction to language technology and its core application areas, together with an evaluation of current language technology support for Welsh.

4

LANGUAGE TECHNOLOGY SUPPORT FOR WELSH

Language technologies are used to develop software systems designed to handle human language and are therefore often called 'human language technologies'. Human language comes in spoken and written forms. While speech is the oldest and in terms of human evolution the most natural form of language communication, complex information and most human knowledge is stored and transmitted through the written word. Speech and text technologies process or produce these different forms of language, using dictionaries, rules of grammar, and semantics. This means that language technology (LT) links language to various forms of knowledge, independently of the media (speech or text) in which it is expressed. Figure 1 illustrates the LT landscape. When we communicate, we combine language with other modes of information and communication media' for example speaking can involve gestures and facial expressions. Digital texts link to pictures and sounds. Movies may contain language in spoken and written form. In other words, speech and text technologies overlap and interact with other multimodal communication and multimedia technologies. In this section, the main application areas of language technology are discussed for Welsh, i.e., language checking, web search, speech interaction, and machine translation. These applications and basic technologies may include, but are not limited to:

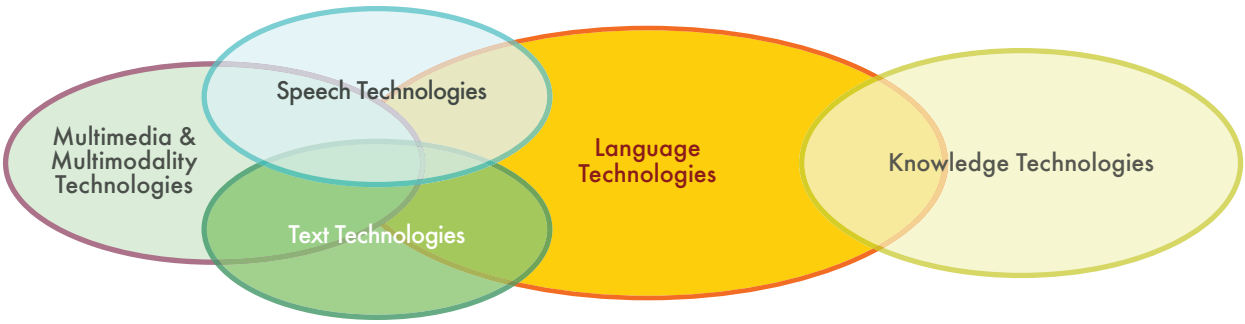
- spelling correction
- authoring support
- computer-assisted language learning
- information retrieval
- information extraction
- text summarisation
- question answering
- speech recognition
- speech synthesis

Language technology is an established area of research with an extensive set of introductory literature. Before discussing the above application areas, the architecture of a typical LT system Application is briefly described.

4.1 APPLICATION ARCHITECTURES

Software applications for language processing typically consist of several components that mirror different aspects of language. While such applications tend to be complex, Figure 2 shows a highly simplified architecture of a typical text processing system. The first three modules handle the structure and meaning of the text input:

1. Pre-processing: cleans the data, analyses or removes formatting, detects the input languages, and so on.
2. Grammatical analysis: finds the verb, its objects, modifiers and other sentence elements; detects the sentence structure
3. Semantic analysis: performs disambiguation (i.e., computes the appropriate meaning of words in a



1: Language technology in context

given context); resolves anaphora (i.e., which pronouns refer to which nouns in the sentence); represents the meaning of the sentence in a machine-readable way.

After analysing the text, task-specific modules can perform other operations, such as automatic summarisation and database look-ups. Note that the architectures of the applications are highly simplified and idealised, to illustrate the complexity of language technology applications in a generally understandable way. In the remainder of this section, an overview of the state of LT research and education for Welsh as it is today is provided, along with a description of past and present Welsh language technology developments. Finally, an estimate of core LT tools and resources for Welsh in terms of various dimensions such as availability, maturity and quality is provided.

4.2 CORE APPLICATION AREAS

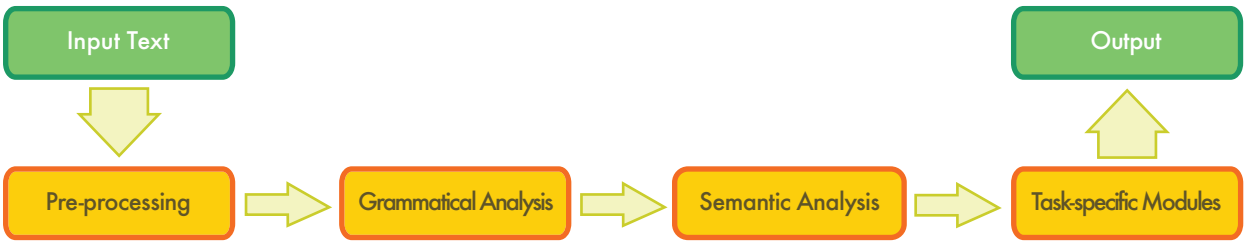
In this section, we focus on the most important LT tools and resources, and provide an overview of LT activities for Welsh.

4.2.1 Language Checking

Anyone who has used a word processor such as Microsoft Word knows that it has a spell checker that highlights spelling mistakes and proposes corrections. Forty years after the first spelling correction programme by Ralph Gorin, language checkers do nowadays do not simply compare the list of extracted words, but have become increasingly sophisticated. Using language-dependent algorithms for **grammatical analysis**, they detect errors related to morphology (e. g., plural formation) as well as syntax-related errors, such as a missing verb or a conflict of verb-subject agreement (e. g., *she *write a letter*). However, most spelling and grammar checkers will not find any errors in the following text [27]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.

Handling these kinds of errors usually requires an analysis of the context. This type of analysis either needs to draw on language-specific **grammars** laboriously coded into the software by experts, or on a statistical language model. In this case, a model calculates the probability of a particular word as it occurs in a specific position. A statistical language model can be automatically created by using a large amount of (annotated) language data (called a **text corpus**). Both of the above approaches have been developed around data from English. Cur-



2: A typical text processing architecture

rently, neither approach can transfer easily to Welsh due to a lack of basic language resources. There are no sufficiently large annotated text corpora to train a statistical model, and there has been insufficient research into the encoding of linguistic knowledge in grammars.

Language checking is not limited to word processors but also applies to authoring systems.

Besides spelling and grammar checkers and authoring support, language checking is also important in the field of computer-assisted language learning. This is an application area which would be of enormous benefit to learners of Welsh and other RMLs.

4.2.2 Spellcheckers, Grammar Checkers, and Computerised Dictionaries

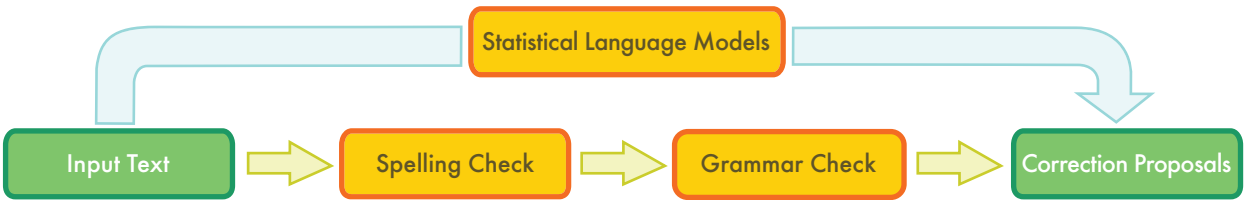
In order to tackle problems regarding content creation in Welsh, several spelling and grammar checkers and computerised dictionaries were commissioned or sponsored by grant in aid. The first was CySill, commissioned from the Departments of Psychology and Linguistics of the University of Wales Bangor. CySill was revolutionary in that it corrected Welsh's initial consonant mutations. This was followed by Cysgair, a computerised dictionary which interfaced with word processing applications, and created by Canolfan Bedwyr of the University of Wales Bangor. Canolfan Bedwyr updated both products in 2004, and included on a single CD of Welsh language computer resources. Amongst the other spellcheckers that already exist in Welsh are the following:

- A free Welsh language spellchecker for Microsoft Office XP, 2003, 2007, 2010, 2013.
- A free Welsh language spellchecker for OpenOffice
- A free Welsh Language Spellchecker for Neo Office (Apple Mac)

LAD (Language Autodetect), a way of detecting languages in a given document, is all the more salient in bilingual settings. How, for example, would a speech recognition system recognise if a speaker changes the language they were speaking at a meeting (either entirely, or by code switching or mixing)? More simply, how does a Welsh speaker construct bilingual English/Welsh documents and have them proofed automatically by the built-in spellchecker without the Welsh being wrongly tagged as mis-spelt English? The Welsh Academy's English-Welsh formerly paper-based dictionary [28] was launched online, a large-scale project funded by the Welsh Language Board and thereafter the Welsh Language Commissioner.

4.2.3 Keyboards, diacritics and fonts

Another element of content creation in Welsh which caused substantial difficulty for computer users in the past was the initial absence of the 'ŵ' and 'ŷ' from the



3: Language checking (top: statistical; bottom: rule-based)

standard character set. Initially, this was dealt with by the creation of specialist Welsh language fonts which mimicked the system fonts included as standard on PCs, but which contained the circumflex on 'w' and 'y'. However, these caused problems on sending a file to a different machine on which those fonts were not installed, the characters being replaced by Icelandic characters. The correct form of both these characters have been included as standard in the Unicode (UTF-8) character set for some time, obviating the need for the purchase or download of specialist fonts for PC users. However, it has not been clear enough how individual end users should access these diacritics in a standardised way, with individuals and different or even the same institutions choosing different keyboard shortcuts, or using character code numbers to insert diacritics into files. This standard diacritic keystroke problem has now been solved, for PC users, in two ways: (1) by the Microsoft UK Extended Keyboard Schema, and (2) by a popular free product called 'To Bach' (Circumflex) manufactured by Draig Technology.

4.2.4 Welsh Language Speech Technology

Speech technology may involve production of a synthetic voice or recognition of a human voice by a given IT system. Such technology is already beginning to permeate our everyday lives (Many call centres have automated their processes using speech synthesis, certain mobile phones which can receive e-mail already offer a synthetic voice facility to read e-mail messages aloud to the recipient). Speech technology can be an asset to any given IT program. It can simplify data access, speed data entry, and allow hands-free control and, significantly, provide biometric passive voice-based authentication for access to secure services such as banking. It obviously also has enormous repercussions for the visually impaired. As speech corpus codification takes substantial effort for any language, it has, historically, been the larger languages which have benefitted from the largest investment, with the RMLs tending to get left behind. As indications show that speech technology, through convergence with other everyday applications, will become a more important part of daily life in the future, it is important that Welsh and other smaller languages secure a strong foothold in this field. Noticing the strategic importance of the field, the Welsh Language Board co-funded, with INTERREG, the WISPR (Welsh and Irish Speech Processing Resources) Project at Canolfan Bedwyr at Bangor University [29]. This project thereafter evolved into the SALT Project [30]. In early 2010 higher quality voices, based on the original basic voices, became available. The initial WISPR project created a basic SAPI compliant (Speech Application Programming Interface) speech synthesis engine for Welsh. Amongst many other uses, speech synthesis engines can be used to convert words from a computer document (e.g. word processor document, web page), or interface into audible speech spoken through the computer sound system. This would be helpful to people who

need or want aural verification of what they are seeing in print. At the main annual Welsh language festival (The National Eisteddfod of Wales) in August 2010, an alpha version of two (one male, one female) high quality synthetic voice was launched. The fact that this was an alpha version which was being launched meant that there have been subsequent waves of improvement (based on feedback via a web interface).

4.2.5 Speech Recognition

Speech recognition, (as defined by the WISPR Project team) or speech-to-text, involves capturing and digitizing the sound waves from a microphone, converting them to basic language units or phonemes, constructing words from phonemes, and contextually analyzing the words to ensure correct spelling for words that sound alike (such as dear and deer). The output is then displayed on the screen as text.

Again, although the technicalities may seem beyond the lay person, the reach and significance of this facility should not be underestimated, as many computer operating systems and handsets *already* offer voice recognition facilities as standard. This is only likely to increase with the passage of time. It is not inconceivable that we will order food, do our banking and a whole host of other services via speech recognition in the future. The Google Translate smartphone app already includes speech recognition, linked with speech synthesis via machine translation. Recognising the importance of developing this field, the Board also gave a grant to create a basic Speech Recognition Engine. The components of this project, and the WISPR speech synthesis project, are freely available on-line. In May 2010, the Independent Review Panel on Bilingual Services of the National Assembly for Wales [31] published its final report. Amongst its many recommendations, technology figured prominently, with a recommendation that more speech technology should be developed in order, in the long term, to create transcripts of meetings semi-automatically. It noted the Assembly's wish to be a leader in the field of bilingual provision, by using technology.

The Welsh Language (Wales) Measure 2011 does not apply to the National Assembly for Wales and the Assembly Commission (the Corporate Body which has responsibility for the provision of property, staff and services to support the Members of the Assembly).

The National Assembly for Wales (Official Languages) Bill was approved by the Assembly on 3 October 2012. The Act's principal objective is to place a statutory duty on the National Assembly for Wales and the Assembly Commission to treat Welsh and English as its official languages and on the basis of equality.

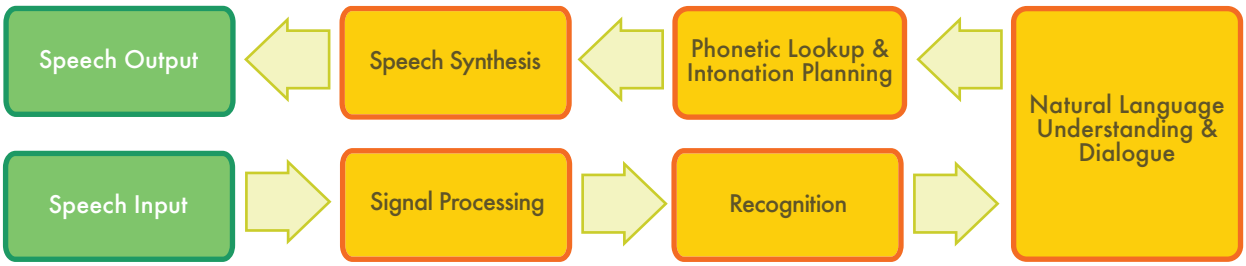
The Act places a duty on the Commission to adopt and publish an Official Languages Scheme [32] specifying the measures it will take in order to comply with its duties as outlined in the Act.

The Scheme was approved by the Assembly on 17 July 2013. It defines the standards and services Members and the public can expect from the Assembly Commission and identifies four key areas for action.

The Assembly Commission will:

- provide innovative, tailored support to enable people to use both languages in the context of Assembly business;
- invest significantly in technology as a way of transforming bilingual service provision whilst also providing value for money;
- develop the Welsh-language skills and confidence of the people who work for the Commission; and
- share experience of delivering bilingual services with other organisations in Wales and legislatures elsewhere and seek to learn from them.

The Assembly's use of translation technology at the moment is limited. Google Translator Toolkit is used



4: Speech-based dialogue system

by their external contractors to produce the Record of Proceedings, supplemented by post-editing and proof-reading to correct and refine the Machine Translation output. Internally the Assembly's Translation and Reporting Service uses Wordfast Translation Memory software, which has resulted in an increase in output.

The Official Languages Scheme commits the Assembly to make best use of technology to translate documents more quickly and efficiently. It has embarked on work to explore the potential benefits of investing in a bespoke machine translation system. As part of this work, they are considering how machine translation can be used not only by the Translation and Reporting Service but other Assembly staff, and Assembly Members and potentially made available to other organisations and institutions beyond the Assembly.

4.2.6 Integration of Machine Translation and Speech Technology

One field which merits consideration in the mid-term is the integration of speech technology with the machine translation technology described above. The ideal scenario would enable two people, one speaking Welsh, the other English, to converse with each other. This would be accomplished by speech recognition, feeding into a machine translation engine, and outputting the relevant translation via speech synthesis. Such integrative technologies are already in production in several institutions, for example The work of Prof Alex Waibel of Carnegie Mellon and Karlsruhe Universities in particular. Such automated technology may soon be implanted in the operating systems we use every day, and therefore must be supported for smaller languages such as Welsh as well. The potential for such integration for Welsh language services, and bilingual meetings, in an age increasingly controlled by IT is obvious.

4.2.7 Translation and Terminology

The Welsh Language Commissioner is responsible for translation policy in Wales and, at the time of writing this white paper, was undertaking an independent review of the profession and its strategic needs. The purpose of this section, then, is to deal with *technology* and translation, and the contribution that technology can make to the Welsh (or any other sociolinguistically similar language) language translation industry.

4.2.8 Computer Assisted Translation

This is the broadest term used to describe an area of language technology applications that automates or assists the act of translating text from one language to another. They are highly effective in improving translation productivity, especially in facilitating very quick translation of repetitive source texts. Some of the most common forms of translation memory are SDLX Trados, Déjà Vu, Wordfast and most recently, Google Transla-

tor Toolkit, Pootle, Transifex and OmegaT. Translation memory software is already used by a range of public institutions in Wales and beyond, amongst them, the Welsh Government, Cardiff University (via a crowdsourcing model) and the National Assembly for Wales itself. One of the main virtues of such technologies is that translators can share each other's translation work, regardless of geographical location, via central databases of translation memories served to individual PCs via corporate networks or even the internet. This means that 100% matches of source text segments can be reused, thereby leading to an increase in the consistency and speed of translation (for information on consistency of terminology, see below). Translation memory applications can also offer translations of partial or 'fuzzy' matches in their databases. This also increases the speed of translation. The main caveat of this approach of course is that the quality of translations shared through centrally stored translation memories depends on the quality and consistency of all those translations fed into them. High-level editorial control is therefore obviously needed on such large-scale projects. One thing that should be consistently emphasised is that human translators are very much needed [33], and that the technology described here aims to improve their productivity and consistency—not to displace them!

4.2.9 Terminology Management

Again, it should be noted at the start of this section that it is not the place of the present document to deal with the general field of standardisation of terminology *per se*, merely to facilitate the diffusion of standardised terminology via IT as an objective of the language normalisation process. Technology already exists, as for the translation memory software described above, to automatically pass standardised lists of terminology to individual end users. This feeds into terminology management applications linked to those translation memory applications. Some such applications also offer disambiguation facilities which provide extra information to end users, e.g. the Welsh version of Mill Street in Aberystwyth (Dan Dre), may not be the same translation as Mill Street in any other town in Wales ('Stryd y Felin' is a grammatically and correct translation, but not used for the street name in Aberystwyth). The word 'access' has many meanings, including as a noun describing the location where one approaches a building, as well as a verb meaning the actual action of gaining entrance to a building, and indeed gaining access to information. 'Mole' can be a small subterranean nocturnal animal, a person who leaks information from an organisation, or an extremely large number normally used in chemistry; all these meanings could be differentiated by means of disambiguation facilities. The larger the size, quality and consistency of a translation memory, the higher the probability that the translator's workload will be achieved in a shorter time. Being aware of the possibilities that TM software offered RMLs, in early 2010. In total around half a million words of bilingual, quality-controlled text were released in TMX format (the industry standard for open exchange of memories), including:

- Human Resources Translation Memory
- Menus Translation Memory
- A translation memory created by aligning the Welsh Language Board's website (containing much public sector vocabulary)

The press release issued with this development called upon the whole public sector to share their translations openly, with the Board offering to be a broker for the data in a translation memory exchange. The main difference between this and other Exchanges is that the data made available would have to be free of charge. All memories are also uploaded to the Google Translator Toolkit described below, and therefore add to the

corpus of example-based translations in Google's Automatic Translation engine, Google Translate. This vision has taken into account wider developments in Web2.0, crowdsourcing and collaborational approaches. In its essence, it believes that nothing but good can come from sharing quality-controlled data between organisations. An over simplistic example often given in presentations is the theoretical case of the 22 local authorities in Wales translating a Council Tax form 22 times, whereas translation memory servers could help automate this substantially. Such sharing and blurring of boundaries between organisations also reflects wider trends in management theory and the agendas of several Government commissioned reports, such as the Gershon Review [34], The Beecham Review [35], and the UK Government Code of Practice [36] on Open Source. It is all-important to mention quality control of data released by public bodies, one element of which would be confidentiality. Translation memories used in public organisations may contain personal data regarding identifiable individuals which must be removed before publication to avoid breach of the UK Data Protection Act. Ways of automating this are already available in such systems. Technology of any sort should not be an end in itself but a strategic policy enabler, in the case of the Welsh language, for linguistic normalisation.

4.2.10 Translation Workflow and Bilingual Document Management

The management of many translation projects occurring at the same time can be greatly aided by IT and computerised workflow systems. These can also monitor availability in the capacity of various external or freelance translators to undertake extra work (such as number of words translatable per day per theme, and the cost rate). The more sophisticated of these 'dashboard' systems can even interface with CAT tools in order to reduce repetitive translation which is contracted out. Such technology can also, of course, manage offices, for example, in institutions such as local authorities who have pools of in-house translators. It would even be possible for external freelance translators to register for such services in order to receive regular work, as well as updates to translation memory and consistent terminology, thereby creating an ever evolving and improving, consistent, quality controlled, corpus of translation memories which can be further shared, further contributing to the normalisation of Welsh. TM systems in the 'cloud' such as the Google Translator Toolkit, further facilitate such collaboration in both terminology and translation memory management.

4.3 TRANSLATION WORKFLOW AND CONTENT MANAGEMENT SYSTEMS

The translation memory systems described above can be be plugged into content/document management systems of all sorts, and linked to workflow solutions for translation units/companies/freelancers and to machine translation systems. In a world where information is constantly changing, it is important that both language versions are simultaneously published and updated. In this vein, it should be recalled that the main tenet of the Welsh Language (Wales) Measure 2011 is that the Welsh language should be treated no less favourably than the English language. Technology has a pivotal role to play in enabling such controlled simultaneous publishing. On, for example, a governmental or local authority website serving the public in Wales, parts of pages or paragraphs may be amended on a daily basis. Where the content management system is managed by non-Welsh speakers, this technology can automatically route amended segments to a translator who will then translate them using translation memory (and terminology management software). Once the transla-

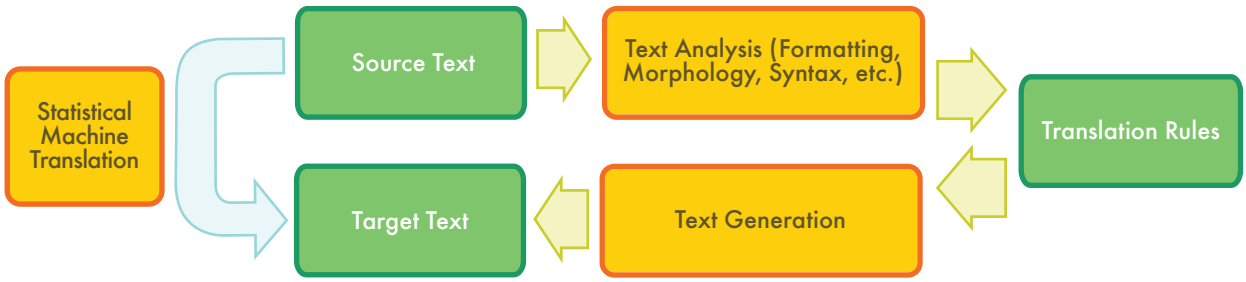
tor has completed translation, the workflow system will automatically route the translation back to the content management system which will publish it simultaneously with the original English version which has been stored awaiting its Welsh language counterpart. This is an important tool for the further normalisation of the Welsh language and will enable end users to have access to up to date English and Welsh text. (One of the results of the *Snapshot Surveys* of websites described above was that the Welsh versions of some public sector websites were not updated as consistently as the English versions, 'Welsh Version to follow' being seen on websites, in some cases, for several years).

4.3.1 Automatic Translation

Automatic translation is a theme often raised when discussing the Welsh language, translation and IT issues. It has benefited from many years of research, and some major advances have been secured. However, such technology without human intervention as 'post editing' cannot yet offer translation quality to match that of human translators. Nevertheless, usage of free on-line systems, particularly Google Translate, released for Welsh in late August 2009, indicates that for certain types of text, it can provide a useful and usable level of translation. Furthermore, in combination with controls on the language used for example in technical authoring, MT can provide an excellent quality of first-draft translation needing little revision, and offering great savings in translation costs. It can also provide gist translation from Welsh, enabling access to the international community, and enabling non-Welsh speaking staff to deal with written correspondence from Welsh speaking colleagues/members of the public.

A 2004 Welsh Language Board feasibility study into machine translation by Professor Harold Somers of Manchester University, and Editor of the *International Journal of Machine Translation*, espoused a tripartite hybrid translation engine: EBMT [Example Based Machine Translation], RBMT [Rules Based Machine Translation], and SBMT [Statistical Based Machine Translation]. This would enable gist translation between Welsh and English and, in the future, integration with other applications (such as the research pane of MS Office and translation memory systems). As stated above, this should be very much considered as an aid to translators, editors et al, rather than a means to supplant them [37]. Moving from translation, to *post-editing*, however, would be a substantial cultural change for many translators and a recognised change management methodology should be used where such a system is to be put in place. Inasmuch, it should be noted that Google Translate, Google's Automatic Translation System, patches into Google Translator Toolkit, SDL Trados Studio 2009, OmegaT and other translation memory systems. This gives the professional translator a full 'cloud-based' dashboard of translation options: automatic [example-based] translation, Translation Memory, and Terminological lists, all with the benefit of real time TM sharing with all other users. This transcends the small corpus statistically based automatic translation engines such as Intertran which have, in the hands of well-meaning but unqualified amateurs created some of the strangest translations ever seen, e.g. 'Staff Entrance' was translated 'correctly' as 'Pastwn Taflu i Berlewyg' [i.e. [large piece of wood [a 'staff']/throw someone into a trance [to 'entrance' them]]. Other examples are too numerous to mention and will, doubtlessly, be familiar to readers in other bilingual areas. Many examples can be found on the Scymraeg picture sharing site [38]. The following machine translation engines are available for Welsh:

- Apertium, developed by the Transducens research group at the Departament de Llenguatges i Sistemes Informàtics of the Universitat d'Alacant in collaboration with Prompsit Language Engineering.



5: Machine translation (left: statistical; right: rule-based)

- Google Translate
- Bangor University's Alpha English-Welsh Machine Translation System
- Intertran

Translation automation could assist RMLs by improving the productivity of minority language translators and by sharing language resources among members of the professional translation community. An increase in translation throughput coupled with an increase in translation quality may lead to a reduction in minority language translation costs, although this is definitely not the sole reason for espousing adoption of such technology. An increase in speed of translation (by, for example, reduction of repetitive translation) could lead to more translations being undertaken. If more minority language translations are being undertaken there is an increased chance that more content will be available in the language, which assists in language 'normalisation'. Making RMLs more visible, especially in modern technologies, is likely to raise the status of the RMLs in the eyes of minority language speakers and possibly increase their desire, and opportunity, to use their RMLs. The Google Translator Toolkit, at the time of writing this white paper, was the only freely available cloud-based TM solution which required no engineering knowledge to use it. Many other TM, both proprietary, and Open Source, exist. The Google Translator Toolkit [39] is a Translation Memory tool that works by the translator uploading a source text, the toolkit then splits the documents into segments, attempts to translate those segments, and then displays different views of the material. On the left hand side of the screen the source document is displayed, on the right hand side the pre-translated document is displayed. A window is also shown where the selected pre-translated segment can be edited, or (in the absence of a pre-translation) translated. At regular intervals, the corpus of translations which are thereby produced following the intervention of the human translator are harvested into the machine translation engine which fuels Google Translate, thereby creating a virtuous circle of translation quality. The larger the corpus of example translations, the higher the statistical probability that the machine translation offered will be of higher quality, needing less post-editing and so on. The possibilities for sharing of translations, in real time, between any number of institutions, without the need for client-side software to be installed, is phenomenal.

Many centres exist, the world over, which investigate the possibility of advanced leveraging of corpora and translation via hybrid machine translation in a workbench environment. Few, however, have taken account of the needs of the human translators themselves who use the technology day in, day out. One such study [40] has done so. Such a workbench could be used by all translators in a given organisation, or a wider network of organisations and would include such facilities, as well as

translation automation, for example, predictive authoring tools, sorting of segments alphabetically, real time progress update towards completion. These linked technologies can further improve the throughput of translators, and therefore improve the linguistic landscape of Wales. We are participating in a quiet technological revolution that will, given necessary investment in suitable language technology components, transform status language planning the world over.

4.3.2 Corpora

Sufficiently sized corpora of written Welsh (which could, for example, contain a large number of electronic versions of printed publications) are a prerequisite for further developments in language technology. Such language technology is the basis for many other Welsh language applications, for example the speech and machine translation technology discussed in this white paper. At the forefront of this field for many years, Canolfan Bedwyr (and others) has developed lemmatisers (which break down given words and tag their grammatical forms), corpora (large databases of written or spoken Welsh), algorithms for sort orders and other language engineering issues. While the majority of these resources *in themselves* will only be of specialist interest, the effect of these very necessary tools is far-reaching and significant, as they feed into other language technology applications. (See also the autoglossing technology referred to above). Below are listed some of the most significant corpora which are available for Welsh.

- *CEG* (Corpws Electroneg o'r Gymraeg/An electronic corpus of the Welsh Language)
- National Foundation for Educational Research, *Ein geiriau ni* (Our Words)
- *Corpws Siarad* (Speech Corpus).

4.4 AVAILABILITY OF TOOLS AND RESOURCES FOR WELSH

Figure 6 provides a rating for language technology support for Welsh. This rating of existing tools and resources was generated from estimates based on a scale from 0 (very low) to 6 (very high) using seven criteria. The key results can be summed up as follows: Although Welsh stands *reasonably* well amongst other RMLs with respect to the most basic language technology tools and resources, such as corpora, inflectional lexicons, tokenisers, interfaces, taggers and lemmatizers, this is no reason to rest on our laurels. Many existing resources lack standardisation so initiatives are needed to standardise the data and interchange formats, for example in terminology management and sharing of translation memories. There exist also individual products with limited functionality in subfields such as speech synthesis, speech recognition and information extraction. At present, only a small number of companies or organisations in Wales are working in the LT area. It is thus extremely important to continue public support for Welsh LT particularly having in mind the enlargement of the Welsh linguistic landscape following the implementation of the Welsh Language (Wales) Measure 2011. It is particularly pleasing to note the recent call for grant applications in this field extended by the Welsh Government based on its 2013 dedicated Welsh Language Technology Strategy Document and Implementation Plan.

4.5 CROSS-LANGUAGE COMPARISON

The current state of LT support varies considerably from one language community to another. In order to compare the situation between languages, this section will present an evaluation based on two sample application areas (machine translation and speech processing)

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	1	1	1	1	1	1	3
Speech Synthesis	1	2	2	2	2	2	3
Grammatical analysis	2	1	2	2	3	2	1
Semantic analysis	2	2	2	2	2	2	2
Text generation	2	2	2	2	2	2	2
Machine translation	3	3	3	2	1	1	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	1	1	2	1	2	2	1
Speech corpora	4	3	4	4	4	4	3
Parallel corpora	3	3	2	3	3	4	3
Lexical resources	3	2	3	2	2	4	4
Grammars	4	3	3	3	3	5	4

6: State of language technology support for Welsh

and one underlying technology (text analysis), as well as basic resources needed for building LT applications. The languages were categorised using the following five point scale:

1. Excellent support
2. Good support
3. Moderate support
4. Fragmentary support
5. Weak or no support

LT support was measured according to the following criteria:

Speech Processing: Quality of existing speech recognition technologies, quality of existing speech synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications.

Machine Translation: Quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications.

Text Analysis: Quality and coverage of existing text analysis technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources and grammars.

Resources: Quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars. Figures 7 to 10 show that Welsh is in the lower cluster for almost all of the tools and resources listed. It compares well with other languages with a small number of speakers,

such as Estonian, Latvian, Lithuanian, Slovak. However, all these languages lag far behind large languages like German and French, for instance. But even LT resources and tools for those languages clearly do not yet reach the quality and coverage of comparable resources and tools for the English language, which is in the lead in almost all LT areas. And there are still plenty of gaps in English language resources with regard to high quality applications.

4.6 CONCLUSIONS

It is clear that historical sociopsychological and cultural barriers to the use of the diglossic 'L' language (in this case Welsh) may still exist on introducing the language to domains such as Language Technology which it previously did not and was not generally expected to inhabit. In all fields, changing habits of language use, which has figured in every Welsh language strategy document in one form or another, takes a long time and language planners need to play a long game. What is clear from the great deal of activity, in both open source and proprietary Welsh language technology is that individuals are amenable to using easily available, high quality Welsh language software when there is awareness and a certain level of demystification of it. The concept of 'Active Offer', first made popular in language services in Canada, and brought to Europe by the *From Act to Action* [41] project, is most salient in this field, as in all other language-related services—if a user is unaware, and not proactively offered a service in a language, how (and indeed why) would the lay person go out of his/her way to find and use it)? The potential for normalisation of Welsh in the field of language technology, and for language technology to normalise Welsh through the orchestrated sharing of big public data is vast.

In this series of white papers, we have made an important effort by assessing the language technology support for 31 European languages, and by providing a high level comparison across these languages. By identifying the gaps, needs and deficits, the European language technology community and its related stakeholders are now in a position to design a large scale research and development programme aimed at building a truly multilingual, technology-enabled communication across Europe. The results of this white paper series show that there is a dramatic difference in language technology support between the various European languages. While there are good quality software and resources available for some languages and application areas, others, usually smaller languages, have substantial gaps. Many languages lack basic technologies for text analysis and the essential resources. Others have basic tools and resources but the implementation of, for example, semantic methods is still far away. Therefore a large-scale effort is needed to attain the ambitious goal of providing high-quality language technology support for all European languages, for example through high quality machine translation. The long-term goal of META-NET is to enable the creation of high-quality language technology for all languages. This requires all stakeholders—in politics, research, business, and society—to unite their efforts. The resulting technology will help tear down existing barriers and build bridges between Europe's languages, paving the way for political and economic unity through cultural diversity.

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch Finnish French German Italian Portuguese Spanish	Basque Bulgarian Catalan Danish Estonian Galician Greek Hungarian Irish Norwegian Polish Serbian Slovak Slovene Swedish	Croatian Icelandic Latvian Lithuanian Maltese Romanian **Welsh**

7: Speech processing: state of language technology support for 31 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	French Spanish	Catalan Dutch German Hungarian Italian Polish Romanian	Basque Bulgarian Croatian Czech Danish Estonian Finnish Galician Greek Icelandic Irish Latvian Lithuanian Maltese Norwegian Portuguese Serbian Slovak Slovene Swedish **Welsh**

8: Machine translation: state of language technology support for 31 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Dutch French German Italian Spanish	Basque Bulgarian Catalan Czech Danish Finnish Galician Greek Hungarian Norwegian Polish Portuguese Romanian Slovak Slovene Swedish	Croatian Estonian Icelandic Irish Latvian Lithuanian Maltese Serbian **Welsh**

9: Text analysis: state of language technology support for 31 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch French German Hungarian Italian Polish Spanish Swedish	Basque Bulgarian Catalan Croatian Danish Estonian Finnish Galician Greek Norwegian Portuguese Romanian Serbian Slovak Slovene	Icelandic Irish Latvian Lithuanian Maltese **Welsh**

10: Speech and text resources: State of support for 31 European languages

ABOUT META-NET

META-NET is a Network of Excellence partially funded by the European Commission. The network currently consists of 54 research centres in 33 European countries. META-NET forges META, the Multilingual Europe Technology Alliance, a growing community of language technology professionals and organisations in Europe. META-NET fosters the technological foundations for a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- grants all Europeans equal access to information and knowledge regardless of their language;
- builds upon and advances functionalities of networked information technology.

The network supports a Europe that unites as a single digital market and information space. It stimulates and promotes multilingual technologies for all European languages. These technologies support automatic translation, content production, information processing and knowledge management for a wide variety of subject domains and applications. They also enable intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots. Launched on 1 February 2010, META-NET has already conducted various activities in its three lines of action META-VISION, META-SHARE and META-RESEARCH.

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. The present White Paper was prepared together with volumes for 29 other languages. The shared technology vision was developed in three sectorial Vision Groups. The META Technology Council was established in order to discuss and to prepare the SRA based on the vision in close interaction with the entire LT community.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, the action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community.

office@meta-net.eu – http://www.meta-net.eu

CYFEIRIADAU REFERENCES

[1] Aljoscha Burchardt, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit, and Martin Volk. *Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012.

[2] Gallup. User language preferences online Flash EB 313: Survey conducted by The Gallup Organization, Hungary upon the request of Directorate-General Information Society and Media Coordinated by Directorate-General Communication. Technical report, European Commission, 2011. http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf.

[3] Ethnologue. Endangered languages, 2013. http://www.ethnologue.com/endangered-languages.

[4] Benedict Anderson. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London, 1983.

[5] European Commission. Multilingualism: an Asset for Europe and a Shared Commitment. Technical report, European Commission, Brussels, 2008. http://ec.europa.eu/languages/pdf/comm2008_en.pdf.

[6] UNESCO. Intersectoral mid-term strategy on languages and multilingualism. Technical report, UNESCO, Paris, 2007. http://unesdoc.unesco.org/images/0015/001503/150335e.pdf.

[7] A. Rinsche and N Portera-Zanotti. Studies on Translation and Multilingualism: The size of the language industry in the EU. Technical report, European Commission, Brussels, 2009. http://bookshop.europa.eu/is-bin/INTERSHOP.enfinity/WFS/EU-Bookshop-Site/en_GB/-/EUR/ViewPublication-Start?PublicationKey=HC8009985.

[8] Ysgol Gwyddorau Cymdeithasol Caerdydd/Cardiff School of Social Sciences School. Cardiff Online Social Media Observatory (COSMOS): Social Media and Data Mining, 2013. http://www.cardiff.ac.uk/socsi/research/researchgroups/comsc-socsi/projects.html.

[9] Llywodraeth Cymru/Welsh Government. Cyfrifiad 2011: Data Iaith Gymraeg ar gyfer Ardaloedd Bach/2011 Census: Welsh Language Data for Small Areas, 2013. http://wales.gov.uk/docs/statistics/2013/130130-2011-census-welsh-language-data-small-areas-cy.pdf, http://wales.gov.uk/docs/statistics/2013/130130-2011-census-welsh-language-data-small-areas-en.pdf.

[10] Comisiynydd y Gymraeg/Welsh Language Commissioner. Cyfrifiad 2011: canlyniadau yn ôl oed/2011 Census: results by age. Technical report, Comisiynydd y Gymraeg/Welsh Language Commissioner, Caerdydd/Cardiff, 2013. http://www.comisiynyddygymraeg.org/Cymraeg/Cymorth/dataacystadegau/Pages/Cyfrifiad2011canlyniadauynoloed.aspx, http://www.comisiynyddygymraeg.org/English/Assistance/Dataandstatisitcs/Pages/2011Censusresultsbyage.aspx.

[11] G Henry. Numbers of children speaking Welsh more than double those of working age or pensioners, 2013. http://www.walesonline.co.uk/news/wales-news/numbers-children-speaking-welsh-more-3864527.

[12] Bwrdd yr Iaith Gymraeg/Welsh Language Board. Amcangyfrif y nifer o siaradwyr Cymraeg yn Lloegr/Estimation of the number of Welsh-speakers in England. Technical report, Bwrdd yr Iaith Gymraeg/Welsh Language Board, Caerdydd/Cardiff, 2007. http://www.webarchive.org.uk/wayback/archive/20120330035632/http://www.byig-wlb.org.uk/english/publications/pages/publicationitem.aspx?puburl=/English/publications/Publications/4844.pdf, http://www.webarchive.org.uk/wayback/archive/20120330020357/http://www.byig-wlb.org.uk/cymraeg/cyhoeddiadau/pages/publicationitem.aspx?puburl=/Cymraeg/cyhoeddiadau/Cyhoeddiadau/4843.pdf.

[13] R. Bandura. *Self-efficacy: the Exercise of Control*. Freeman, New York, 1997.

[14] R.M Jones. Cymraeg Iach. *Y Traethodydd*, CXXXIX:590–593, 1984.

[15] Cymdeithas yr Iaith Gymraeg/The Welsh Language Society. Gwefan/Website, 2013. http://www.cymdeithas.com.

[16] Llywodraeth Cymru/Welsh Government. Strategaeth y Gymraeg: Cynllun Gweithredu 2013 i 2014 – Iaith fyw: iaith byw/Welsh language strategy: Action Plan 2013 to 2014 – A living language: a language for living. Technical report, Llywodraeth Cymru/Welsh Government, Caerdydd/Cardiff, 2013. http://wales.gov.uk/topics/welshlanguage/publications/welsh-language-strategy-action-plan-2013-14/?skip=1&lang=cy, http://wales.gov.uk/topics/welshlanguage/publications/welsh-language-strategy-action-plan-2013-14/?skip=1&lang=en.

[17] Llywodraeth Cymru/Welsh Government. Cynllun Gweithredu Technoleg a Chyfryngau Digidol Cymraeg/Welsh Language Technology and Digital Media Action Plan. Technical report, Llywodraeth Cymru/Welsh Government, Caerdydd/Cardiff, 2013. http://wales.gov.uk/docs/dcells/publications/230513-action-plan-cy.pdf, http://wales.gov.uk/docs/dcells/publications/230513-action-plan-en.pdf.

[18] Richard Sheppard, Jeremy Evas, and Canolfan Bedwyr. *Canllawiau a Safonau Meddalwedd/Bilingual Software Guidelines and Standards*. Bwrdd yr Iaith Gymraeg/Welsh Language Board, Caerdydd/Cardiff, Ebrill/April 2006. http://www.webarchive.org.uk/wayback/archive/20120330005938/http://www.byig-wlb.org.uk/cymraeg/cyhoeddiadau/Pages/PublicationItem.aspx?puburl=/Cymraeg/cyhoeddiadau/Cyhoeddiadau/3962.pdf, http://www.webarchive.org.uk/wayback/archive/20120330013521/http://www.byig-wlb.org.uk/english/publications/Pages/PublicationItem.aspx?puburl=/English/publications/Publications/3963.pdf.

[19] Bwrdd yr Iaith Gymraeg/Welsh Language Board. Technoleg Gwybodaeth a'r Gymraeg: Dogfen Strategaeth/Information Technology and Welsh: A Strategy Document. Technical report, Bwrdd yr Iaith Gymraeg/Welsh Language Board, Caerdydd/Cardiff, 2006. http://www.webarchive.org.uk/wayback/archive/20120330022112/http://www.byig-wlb.org.uk/cymraeg/cyhoeddiadau/pages/publicationitem.aspx?puburl=/Cymraeg/cyhoeddiadau/Cyhoeddiadau/3964.pdf, http://www.webarchive.org.uk/wayback/archive/20120330041214/http://www.byig-wlb.org.uk/english/publications/pages/publicationitem.aspx?puburl=/English/publications/Publications/3965.pdf.

[20] Jeremy Evas. Ciparolwg: Safleoedd gwe cyrff sydd â Chynllun Iaith statudol. Technical report, Bwrdd yr Iaith Gymraeg, Caerdydd, 2001. http://orca.cf.ac.uk/43867/.

[21] Rhiannon Gomer. Ciparolwg Gwefannau (2003): Ciparolwg Annibynnol ar ddarpariaeth ddwyieithog gwefannau cyrff sydd â Chynllun Iaith Statudol – Snapshot Survey of Websites (2003): An independent snapshot survey on the bilingual provision of the websites of bodies which have statutory language schemes. Technical report, Bwrdd yr Iaith Gymraeg/Welsh Language Board, Caerdydd/Cardiff, 2003. http://www.webarchive.org.uk/wayback/archive/20120330022442/http://www.byig-wlb.org.uk/cymraeg/cyhoeddiadau/pages/publicationitem.aspx?puburl=/Cymraeg/cyhoeddiadau/Cyhoeddiadau/409.pdf, http://www.webarchive.org.uk/wayback/archive/20120330041238/http://www.byig-wlb.org.uk/english/publications/pages/publicationitem.aspx?puburl=/English/publications/Publications/410.pdf.

[22] Bwrdd yr Iaith Gymraeg/Welsh Language Board. Cynllun Achredu ar gyfer Meddalwedd Dwyieithog/Accreditation Scheme for Bilingual Software. Technical report, Bwrdd yr Iaith Gymraeg/Welsh Language Board, Caerdydd/Cardiff, 2009. http://orca.cf.ac.uk/43852/.

[23] Canolfan Bedwyr. ECDL (European Computer Driving Licence/Trwydded Yrru Gyfrifiadurol Ewropeaidd, 2013. http://www.bangor.ac.uk/itservices/ecdl/.

[24] Llywodraeth Cymru/Welsh Government. Statistical bulletin: National survey for wales – internet results. Technical report, Caerdydd/Cardiff, Mawrth/March 2012. http://wales.gov.uk/docs/statistics/2012/121219sb1202012en.pdf.

[25] Office for National Statistics. Internet access quarterly update, 2013. http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-303599.

[26] Wikimedia. List of Wikipedias, 2013. http://meta.m.wikimedia.org/wiki/List_of_Wikipedias.

[27] Jerrold H. Zar. Candidate for a pullet surprise. *Journal of Irreproducible Results*, page 13, 1994.

[28] Bruce Griffiths and Dafydd Glyn Jones. Geiriadur yr Academi/The Welsh Academy English-Welsh Dictionary, 2012. http://techiaith.bangor.ac.uk/GeiriadurAcademi/?lang=en.

[29] Canolfan Bedwyr. WISPR – Welsh and Irish Speech Processing Resources, 2013. http://www.bangor.ac.uk/canolfanbedwyr/wispr.php.cy.

[30] SALT. SALT Cymru, Technoleg Iaith a Lleferydd, Speech and Language Technology, 2012. http://www.saltcymru.org/wordpress/.

[31] Cynulliad Cenedlaethol Cymru/National Assembly for Wales. Adolygiad o Wasanaethau Dwyieithog, Adroddiad Terfynol /The Independent Review Panel on Bilingual Services: Final Report, 2010. http://www.assemblywales.org/cy/review-of-bilingual-services-report-english.pdf, http://www.assemblywales.org/review-of-bilingual-services-report-english.pdf.

[32] Cynulliad Cenedlaethol Cymru/National Assembly for Wales. Y Cynllun Ieithoedd Swyddogol – Official Languages Scheme, 2013. http://www.assemblywales.org/cy/bus-home/bus-business-fourth-assembly-laid-docs/gen-ld9401-e.pdf?langoption=3&ttl=GEN-LD9401%20-%20Comisiwn%20y%20Cynulliad%3A%20Y%20Cynllun%20Ieithoedd%20Swyddogol.

[33] A. Way. Who says machine translation will replace translators?, 2013. http://www.lingo24.com/blogs/company/who-says-machine-translation-will-replace-translators.html.

[34] Sir Peter Gershon. Releasing Resources to the Front Line: An Independent Review of Public Sector Efficiency. Technical report, Her Majesty's Stationery Office, London, 2004. http://webarchive.nationalarchives.gov.uk/20130129110402/http://www.hm-treasury.gov.uk/d/efficiency_review120704.pdf.

[35] Jeremy Beecham. Creu'r Cysylltiadau – Cyflawni Ar Draws Ffiniau: Gweddnewid Gwasanaethau Cyhoeddus yng Nghymru/Making the Connections – Delivering Beyond Boundaries: Transforming Public Services in Wales. Technical report, Llywodraeth Cynulliad Cymru/Welsh Assembly Government, Caerdydd/Cardiff, 2006. http://webarchive.nationalarchives.gov.uk/20060715141954/http://new.wales.gov.uk/topics/improvingservices/strategypolicy/delivering/?lang=cy, http://webarchive.nationalarchives.gov.uk/20060715141954/new.wales.gov.uk/topics/improvingservices/strategypolicy/delivering/?lang=en.

[36] HM Government. Open Source, Open Standards and Re-use: Government Action Plan, 2010. https://www.gov.uk/government/publications/open-source-open-standards-and-re-use-government-action-plan.

[37] Piet Verleysen. MT at Work Conference: by Practitioners for Practitioners. *Language and Translation No 6: Machine Translation*, 6:6–9, 2013. http://ec.europa.eu/dgs/translation/publications/magazines/languagestranslation/documents/issue_06_en.pdf.

[38] Sgymraeg. Sgymraeg, 2013. http://www.flickr.com/photos/tags/sgymraeg/.

[39] Google. Google Translator Toolkit, 2013. http://translate.google.com/toolkit.

[40] CASMACAT. CASMACAT: Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation – First Year Report. Technical report, CASMACAT, 2012. http://www.casmacat.eu/index.php?n=Main.FirstYear.

[41] C. H. Williams, S Sandberg, and P. Ó'Flatharta. From Act to Action: Implementing Language Legislation in Finland, Ireland and Wales. Technical report, Dublin City University, Dublin, 2013.

[42] Cynulliad Cenedlaethol Cymru/National Assembly for Wales. Mesur y Gymraeg (Cymru) 2011/Welsh Language (Wales) Measure 2011, 2011. http://www.legislation.gov.uk/mwa/2011/1/contents/enacted/welsh.

[43] Jeremy Evas. Snapshot survey: websites of organisations complying with statutory Welsh Language Schemes. Technical report, Welsh Language Board, Cardiff, 2001. http://orca.cf.ac.uk/43868/.

AELODAU META-NET META-NET MEMBERS

Gwlad Belg	Belgium	Computational Linguistics and Psycholinguistics Research Centre, University of Antwerp: Walter Daelemans
		Centre for Proc. Speech and Images, University of Leuven: Dirk van Compernolle
Bwlgaria	Bulgaria	Institute for Bulgarian Language, Bulgarian Academy of Sciences: Svetla Koeva
Cyprus	Cyprus	Language Centre, School of Humanities: Jack Burston
Croatia	Croatia	Institute of Linguistics, Faculty of Humanities and Social Science, University of Zagreb: Marko Tadić
Denmarc	Denmark	Centre for Language Technology, University of Copenhagen: Bolette Sandford Pedersen, Bente Maegaard
Estonia	Estonia	Institute of Computer Science, University of Tartu: Tiit Roosmaa, Kadri Vider
Swistir	Switzerland	Idiap Research Institute: Hervé Bourlard
Iwerddon	Ireland	School of Computing, Dublin City University: Josef van Genabith
Y Ffindir	Finland	Comp. Cognitive Systems Research Group, Aalto University: Timo Honkela
		Department of Modern Languages, University of Helsinki: Kimmo Koskenniemi, Krister Lindén
Ffrainc	France	Centre National de la Recherche Scientifique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur and Institute for Multilingual and Multimedia Information: Joseph Mariani
		Evaluations and Language Resources Distribution Agency: Khalid Choukri
Yr Almaen	Germany	Language Technology Lab, DFKI: Hans Uszkoreit, Georg Rehm
		Human Language Technology and Pattern Recognition, RWTH Aachen University: Hermann Ney
		Department of Computational Linguistics, Saarland University: Manfred Pinkal
Gwlad Groeg	Greece	R.C. "Athena", Institute for Language and Speech Processing: Stelios Piperidis
Yr Eidal	Italy	Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli": Nicoletta Calzolari
		Human Language Tech. Research Unit, Fondazione Bruno Kessler: Bernardo Magnini
Norwy	Norway	Department of Linguistic, University of Bergen: Koenraad De Smedt

		Department of Informatics, Language Technology Group, University of Oslo: Stephan Oepen
Gwlad yr Iâ	Iceland	School of Humanities, University of Iceland: Eiríkur Rögnvaldsson
Yr Iseldiroedd	Netherlands	Utrecht Institute of Linguistics, Utrecht University: Jan Odijk
		Computational Linguistics, University of Groningen: Gertjan van Noord
Latfia	Latvia	Tilde: Andrejs Vasiļjevs
		Inst. of Mathematics and Comp. Science, University of Latvia: Inguna Skadiņa
Lithwania	Lithuania	Institute of the Lithuanian Language: Jolanta Zabarskaitė
Lwcsembwrg	Luxembourg	Arax Ltd.: Vartkes Goetcherian
Malta	Malta	Department Intelligent Computer Systems, University of Malta: Mike Rosner
Awstria	Austria	Zentrum für Translationswissenschaft, Universität Wien: Gerhard Budin
Gwlad Pwyl	Poland	Institute of Computer Science, Polish Academy of Sciences: Adam Przepiórkowski, Maciej Ogrodniczuk
		University of Łódź: Barbara Lewandowska-Tomaszczyk, Piotr Pęzik
		Department of Computer Linguistics and Artificial Intelligence, Adam Mickiewicz University: Zygmunt Vetulani
Portiwgal	Portugal	University of Lisbon: António Branco, Amália Mendes
		Spoken Language Systems Laboratory, Institute for Systems Engineering and Computers: Isabel Trancoso
Gweriniaeth Tsiec	Czech Republic	Institute of Formal and Applied Linguistics, Charles University in Prague: Jan Hajič
Y DU	UK	School of Computer Science, University of Manchester: Sophia Ananiadou
		Institute for Language, Cognition and Computation, Center for Speech Technology Research, University of Edinburgh: Steve Renals
		Research Institute of Informatics and Language Processing, University of Wolverhampton: Ruslan Mitkov
Rwmania	Romania	Research Inst. for Artificial Intelligence, Romanian Academy of Sciences: Dan Tufiș
		Faculty of Computer Science, University Alexandru Ioan Cuza of Iași: Dan Cristea
Serbia	Serbia	University of Belgrade, Faculty of Mathematics: Duško Vitas, Cvetana Krstev, Ivan Obradović
		Pupin Institute: Sanja Vranes
Slofenia	Slovenia	Jožef Stefan Institute: Marko Grobelnik
Slofacia	Slovakia	Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences: Radovan Garabík
Sbaen	Spain	Barcelona Media: Toni Badia, Maite Melero
		Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: Núria Bel

		Aholab Signal Processing Laboratory, University of the Basque Country: Inma Hernaez Rioja
		Center for Language and Speech Technologies and Applications, Universitat Politècnica de Catalunya: Asunción Moreno
		Department of Signal Processing and Communications, University of Vigo: Carmen García Mateo
Sweden	Sweden	Department of Swedish, University of Gothenburg: Lars Borin
Hwngari	Hungary	Research Institute for Linguistics, Hungarian Academy of Sciences: Tamás Váradi
		Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics: Géza Németh, Gábor Olaszy

Fe wnaeth oddeutu 100 o arbenigwyr – cynrychiolwyr gwledydd ac ieithoedd META-NET – drafod a rhoi terfynol wedd ar ganlyniadau a negeseuon allweddol y gyfres o bapurau gwyn mewn cyfarfod META-NET yn Berlin, yr Almaen ar Hydref 21/22, 2011. — About 100 language technology experts – representatives of the countries and languages represented in META-NET – discussed and finalised the key results and messages of the White Paper Series at a META-NET meeting in Berlin, Germany, on October 21/22, 2011.

CYFRES PAPURAU GWYN META-NET

THE META-NET WHITE PAPER SERIES

Basgeg	Basque	euskara
Saesneg	English	English
Bokmål Norwyeg	Norwegian Bokmål	bokmål
Bwlgareg	Bulgarian	български
Catalaneg	Catalan	català
Croateg	Croatian	hrvatski
Daneg	Danish	dansk
Estoneg	Estonian	eesti
Ffinneg	Finnish	suomi
Ffrangeg	French	français
Galisieg	Galician	galego
Gwyddeleg	Irish	Gaeilge
Almaeneg	German	Deutsch
Groeg	Greek	ελληνικά
Eidaleg	Italian	italiano
Islandeg	Icelandic	íslenska
Latfieg	Latvian	latviešu valoda
Lithwanieg	Lithuanian	lietuviu kalba
Malteg	Maltese	Malti
Nynorsk Norwyeg	Norwegian Nynorsk	nynorsk
Iseldireg	Dutch	Nederlands
Pwyleg	Polish	polski
Portiwgaleg	Portuguese	português
Rwmaneg	Romanian	româna
Serbeg	Serbian	српски
Tsieceg	Czech	ceština
Slovaceg	Slovak	slovencina
Sloveneg	Slovene	slovenščina
Sbaeneg	Spanish	español
Swedeg	Swedish	svenska
Cymraeg	Welsh	Cymraeg
Hwngareg	Hungarian	magyar

In everyday communication, Europe's citizens, business partners and politicians are inevitably confronted with language barriers. Language technology has the potential to overcome these barriers and to provide innovative interfaces to technologies and knowledge. This white paper presents the state of language technology support for the Welsh language. It is part of a series that analyzes the available language resources and technologies for 31 European languages. The analysis was carried out by META-NET, a Network of Excellence funded by the European Commission. META-NET consists of 54 research centres in 33 countries, who cooperate with stakeholders from economy, government agencies, research organisations, non-governmental organisations, language communities and European universities. META-NET's vision is high-quality language technology for all European languages.

Wrth gyfathrebu bob dydd, mae'n anochel bod dinasyddion, partneriaid busnes a gwleidyddion Ewrop yn wynebu rhwystrau ieithyddol. Mae gan dechnoleg iaith y potensial i oresgyn y rhwystrau hyn ac i fod yn rhyngwyneb arloesol i dechnolegau a gwybodaeth. Mae'r gyfres hon o bapurau gwyn yn cyflwyno cyflwr yr adnoddau a thechnoleg iaith sydd ohoni i 31 o ieithoedd Ewrop. Gwnaed y dadansoddiad gan META-NET, Rhwydwaith o Ragoriaeth a ariannwyd gan y Comisiwn Ewropeaidd. Mae META-NET yn cynnwys 54 Canolfan Ymchwil mewn 33 gwlad, sy'n cydweithredu gyda rhanddeiliaid o'r economi, asiantaethau llywodraethol, sefydliadau ymchwil, sefydliadau anllywodraethol, cymunedau ieithyddol a Phrifysgolion yn Ewrop. Gweledigaeth META-NET yw technoleg iaith uchel ei hansawdd i bob un o ieithoedd Ewrop.

"All too often efforts to safeguard and sustain the Welsh language are concerned with trying to preserve the past. This paper looks to the future and demonstrates how technology can assist the language and its speakers occupy their rightful space in the digital world. It draws attention to how much has yet to be done if Welsh is to take full advantage of what technology has to offer, since the language receives weak or no support at best in each of the measured categories. Essential reading for policy makers and language planners."
– Rhodri Williams (Director, Wales, OFCOM)

"In an increasingly connected world, much power vests in those who own, control or shape the technology that connects us. In this context, most languages are minority languages, struggling in the first place for recognition, and then for resources and tools to facilitate their free use. In reviewing these developments for Welsh, and putting them in a global context, this META-NET white paper offers an intelligent and articulate review of the issues faced, and strategies available to all languages in this position, not only those traditionally considered to be in the category of "minority". It is required reading for professional language planners and others concerned to preserve language rights and promote linguistic diversity."
– Emyr Lewis (Member, Committee of Experts of the European Charter for Regional and Minority Languages 2001-2013)