Gwnewch Ramadegau Wici!

Mélanie Jouitteau, Loïc Grobol

▶ **To cite this version:**

Mélanie Jouitteau, Loïc Grobol. Gwnewch Ramadegau Wici!. Iaith a Thechnoleg yng Nghymru, II, Gareth Watkins, 2024, 978-1 84220-208-1. hal-04793349

HAL Id: hal-04793349
https://hal.science/hal-04793349v1

Submitted on 20 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

PRIFYSGOL BANGOR / BANGOR UNIVERSITY

Iaith a Thechnoleg yng Nghymru: Cyfrol II

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

Cyhoeddwyd: 01/11/2024

PDF y cyhoeddwr, a elwir hefyd yn Fersiwn o'r cofnod

Cyswllt i'r cyhoeddiad / Link to publication

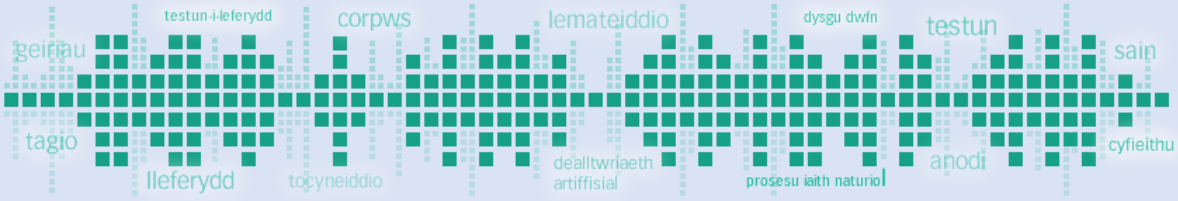
Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):
Watkins, G. (Gol.), Prys, D., Prys, G., Jones, D., Ghazzali, S., Vangberg, P., Farhat, L., Cooper, S., Williams, M., Gruffydd, I., Jouitteau , M., Grobol, L., Morris , J., Ezeani , I., Young , K., Davies, L., El-Haj , M., Knight , D., Jarvis , C., & Barnes, E. (2024). *Iaith a Thechnoleg yng Nghymru: Cyfrol II*. Prifysgol Cymru Bangor.

Hawliau Cyffredinol / General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

20. Nov. 2024



Iaith a Thechnoleg yng Nghymru: Cyfrol II

Golygydd: Gareth Watkins

Cyhoeddwyd yr e-lyfr hwn gyntaf yn 2024 gan

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

www.bangor.ac.uk/cy

Rhif Llyfr Rhyngwladol (e-lyfr):

ISBN: 978-1 84220-208-1.

Mae'r testun wedi'i ryddhau dan y drwydded Creative Commons BY 4.0 https://creativecommons.org/licenses/by/4.0/, sy'n eich caniatáu i'w ailddefnyddio a'i newid mewn unrhyw ffordd os ydych yn rhoi cydnabyddiaeth briodol. Gweler testun y drwydded https://creativecommons.org/licenses/by/4.0/ am ragor o fanylion.

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

Gwnewch Ramadegau Wici!

MÉLANIE JOUITTEAU

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

LOÏC GROBOL

Université Paris Nanterre

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

Allweddeiriau: Ramadegau Wici, NLP, Llydaweg, Corpws

1 CYFLWYNIAD

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

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

2 AMRYWIAETH IEITHYDDOL DRWY DDYLUNIAD

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

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

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

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

2.1 Amrywiaeth dafodieithol a hanesyddol

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

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

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

2.2 Maint

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

3 FFYNHONNELL DATA AR GYFER NLP

3.1 Anodiadau morffogystrawennol

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

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

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

3.2 Data cyfochrog

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

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

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

3.3 Amcangyfrifon cost

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

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

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

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

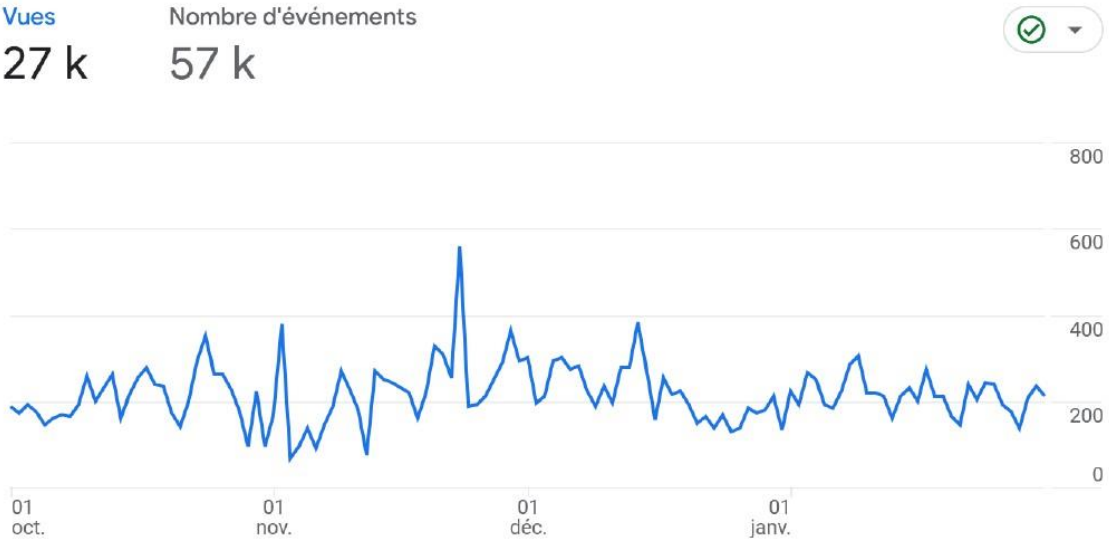
Argymhellir datblygu gramadegau wici yn arbennig ar gyfer adeiladu adnoddau projectau peilot i ieithoedd gyda chorpora cyfyngedig, oherwydd hyd yn oed lle mae gweithredwyr technoleg am y tro yn methu darparu offer gorffenedig ar gyfer siaradwyr, bydd y buddsoddiad yn parhau yn werthfawr ar gyfer y gymuned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

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

4 YMGYSYLLTU CYMDEITHASOL MEWN IEITHOEDD LLEIFAFRIFOL

4.1 Ymgysylltiad cyhoeddus

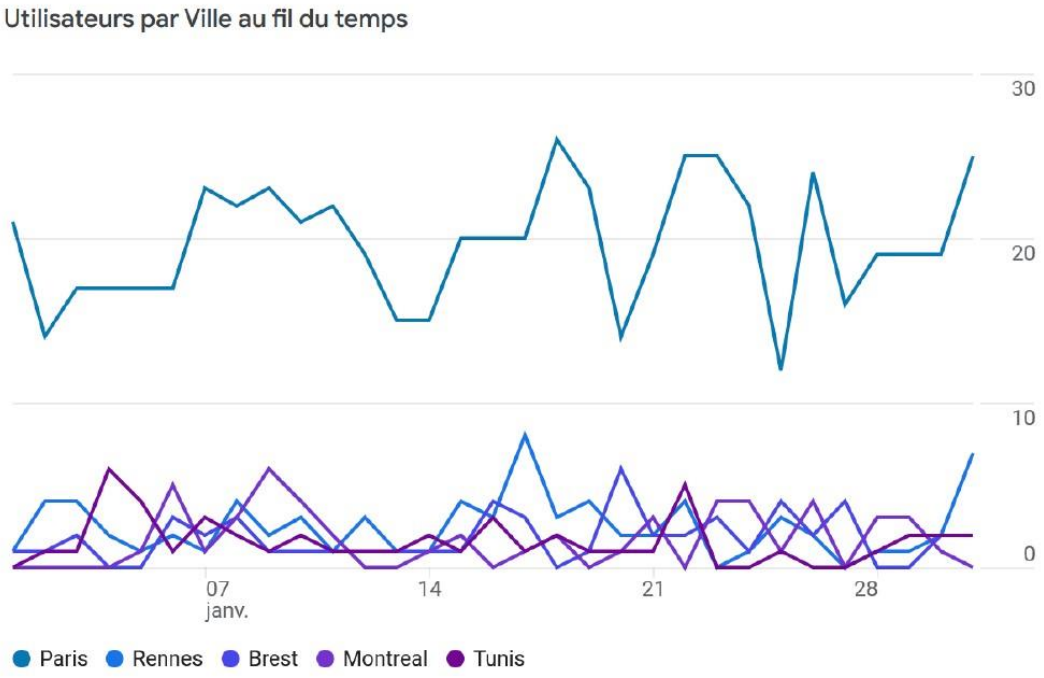
Mae offer ystadegol mewnol, yn ogystal â systemau dadansoddi allanol yn rhoi gwybodaeth fanwl am ddefnydd y wefan, gan dracio (heb fanylion defnyddwyr) y 100 a mwy o ymweliadau dynol dyddiol ag ARBRES. Mae'r graff yn Ffigur 1 yn dangos ystadegau ymweliadau o bob man drwy'r byd o Hydref 2023 hyd at ddiwedd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



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

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

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



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

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

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

4.2 Arfogi'r rhyngwyneb rhwng gwyddoniaeth a chymdeithas

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

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

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

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

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

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

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

CYFEIRIADAU

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

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

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

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

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

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

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