Economic and Social Research Council

Arts and Humanities Research Council

www.corcencc.cymru

CorCenCC

Corpws Cenedlaethol Cymraeg Cyfoes

National Corpus of Contemporary Welsh

Cardiff University / Prifysgol Caerdydd

Swansea University / Prifysgol Abertawe

Prifysgol Bangor University

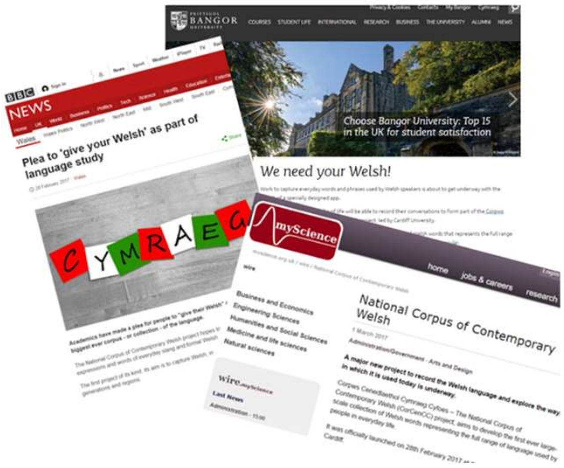
Lancaster University

Trosolwg

Mae CorCenCC yn gasgliad hygyrch o samplau o Gymraeg, wedi'u casglu o gyfathrebu bywyd go iawn, ac wedi'u cadw fel cronfa ddata electronig ('corpws'). Gallwch chwilio CorCenCC i ddarganfod mwy am sut mae pobl wir yn defnyddio'r Gymraeg, er enghraifft, pa mor aml y defnyddir gair penodol, neu beth yw'r geiriau mwyaf aml eu defnydd mewn mathau penodol o gyfathrebu (neu ar draws y corpws cyfan).

Mae CorCenCC yn cynnwys dros 11 miliwn o eiriau o ffynonellau Cymraeg ysgrifenedig, llafar ac electronig (testunau digidol, ar-lein), wedi'u cymryd o amrywiaeth o genres, amrywiadau iaith (rhanbarthol a chymdeithasol) a chyd-destun (gw. tudalen 2 am fwy o fanylion). Mae pob gair mewn corpws yn cael ei 'dagio', er enghraifft gyda gwybodaeth ramadegol (h.y. rhannau ymadrodd – enw, berf, ac ati) a gwybodaeth semantig (yn ymwneud â themâu a phynciau). Rhoddir gwybodaeth am darddiad pob darn o iaith (e.e. math o destun, lleoliad siaradwyr). Golyga hyn fod CorCenCC yn offeryn electronig gwerthfawr sy'n gadael i ni archwilio a deall ein hiaith yn well.

Mae'r llyfryn hwn yn dangos pwy allai elwa o ddefnyddio CorCenCC a sut mae modd defnyddio'r corpws, yn ogystal â phroffilio ambell brosiect lloeren diweddar sydd wedi dod yn sgil datblygu'r adnodd cenedlaethol hwn.



Cynnwys

|**Tud. 1:** Trosolwg

|**Tud. 2:** Manylion Data

|**Tud. 3:** Nodweddion: v2.0

|**Tud. 4-7:** Dysgwyr ac Athrawon

|**Tud. 8:** Ymchwilwyr

|**Tud. 9:** Geiriadurwyr a Chyfieithwyr

|**Tud. 10:** Prosiectau Lloeren

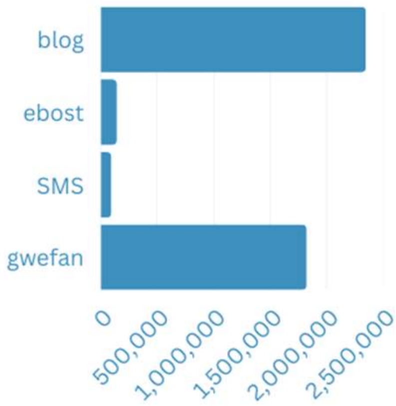
|**Tud. 11:** Darganfod Mwy

Manylion Data

|3.9m o eiriau o destunau **ysgrifenedig**

|4.4m o eiriau o **iaith electronig** (testunau digidol)

|2.86m o eiriau o'r iaith **lafar**



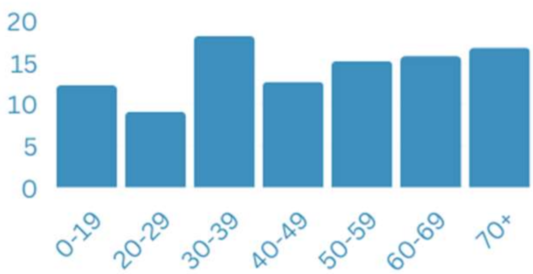
|cyfri'r geiriau mewn testunau **Iaith electronig**

|geiriau o destunau **ysgrifenedig**

|o'r holl lyfrau **ysgrifenedig** mae:

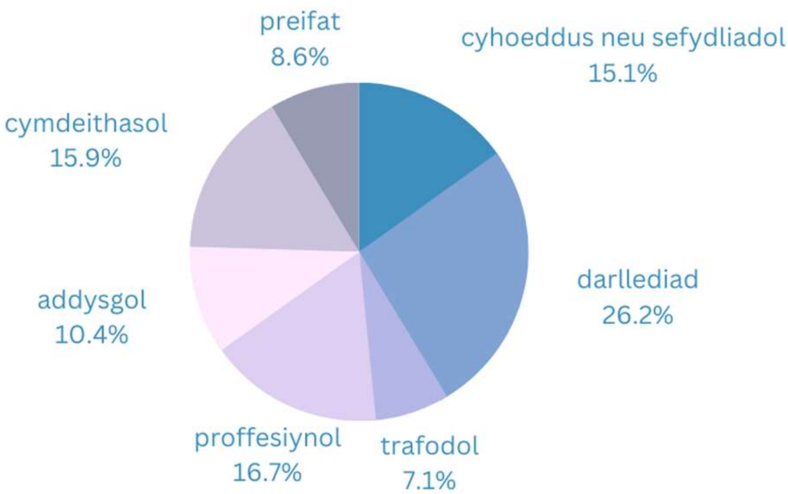
> 22% ar gyfer oedolion
> 11.25% ar gyfer plant
> 4% ar gyfer dysgwyr sy'n oedolion
> 3.5% ar gyfer dysgwyr ifanc

|oedran cyfranwyr data **llafar** (fesul %)



|cyd-destunau recordio data **llafar**

|o gyfranwyr y data **llafar** a nododd eu bod yn fenywod

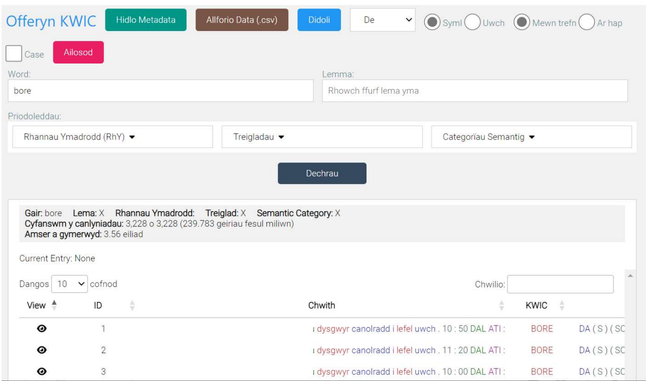


Infographics produced in Canva.com

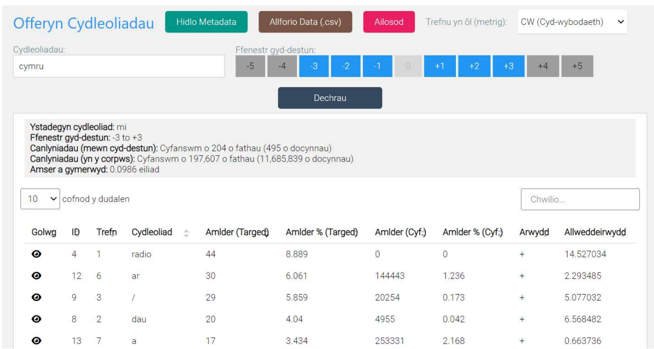
Nodweddion:v2.0

KWIC (Allweddair mewn Cyd-destun)

KWIC yw'r ffordd ragosodedig i archwilio CorCenCC. Gallwch deipio unrhyw air, ymadrodd, categori gramadegol (rhan ymadrodd – RhY) neu semantig ym mlwch chwilio'r offeryn hwn a darganfod pa mor aml mae'n digwydd yn y corpws cyfan, neu mewn is-set benodol o ddata'r corpws. Gallwch hidlo trwy gyfyngu eich chwiliad i foddau gwahanol (ysgrifenedig, llafar, iaith electronig), genre(s), cyd-destun(au), yn ôl oedran, rhyw neu leoliad cyfrannwr, er enghraifft. Fe welwch yn KWIC enghreifftiau o'r gair a chwiliwyd mewn cyd-destun, fel mae'n cael ei ddefnyddio. Hefyd, mae'r offeryn KWIC yn dweud wrthoch chi pa mor aml mae'r term rydych yn chwilio amdano yn cael ei ddefnyddio yn y set ddata.



|Offer **KWIC** (uchod) a **Chydleoliad** (isod)



Rhestr Amlder

Gallwch ddarganfod pa mor aml mae gair a/neu lema (h.y. sail gair a'i ffurfdroadau) yn digwydd yn set ddata gyfan CorCenCC, neu mewn is-setiau o'r data (gw. **KWIC**).
Mae modd creu rhestri amlder o gategorïau gramadegol (RhY), mathau o dreigladau a chategorïau semantig hefyd.

N-gramau

Fel gyda'r offeryn rhestr **amlder**, gallwch greu rhestri o glystyrau cyffredin (n-gramau) o eiriau sy'n digwydd yn y corpws, o'r ddau air mwyaf cyffredin sy'n cyd-ddigwydd i batrymau o bum gair.

Cydleoliadau

Mae **cydleoliad** yn cyfeirio at y berthynas rhwng geiriau a'r atyniad cymharol rhyngddynt. Gallwch chwilio am **gydleoliadau** gair penodol (neu dag RhY/semantig) o fewn ffenestr benodol, h.y. hyd at 5 gair ar bob ochr i'r gair chwilio hwnnw.

Allweddair

Mae dadansoddiad **allweddair** yn gadael i chi gymharu is-setiau gwahanol o ddata'r corpws â'i gilydd i benderfynu pa eiriau, ymadroddion, tagiau ac ati, sy'n fwy tebygol o ddigwydd yn fwy aml mewn un is-set o'i chymharu â'r llall.

Dysgwyr ac Athrawon

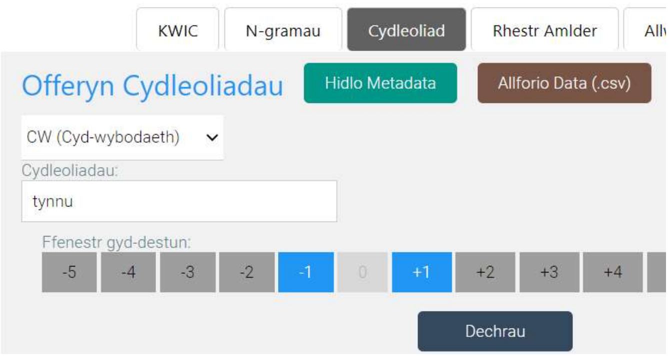
B1 Canolradd: cydleoliadau (tasg enghreifftiol)

NOD: Trwy ddefnyddio’r offeryn dadansoddi **cydleoliad**, bydd myfyrwyr yn dysgu sut i nodi cydleoliadau cyffredin a gysylltir â geiriau chwilio gwahanol sydd o ddiddordeb iddynt (**tynnu** yn yr achos yma). Mae’n ffordd o ymestyn gwybodaeth eirfaol myfyrwyr ar sail y geiriau ddysgon nhw ar lefelau A1 ac A2.

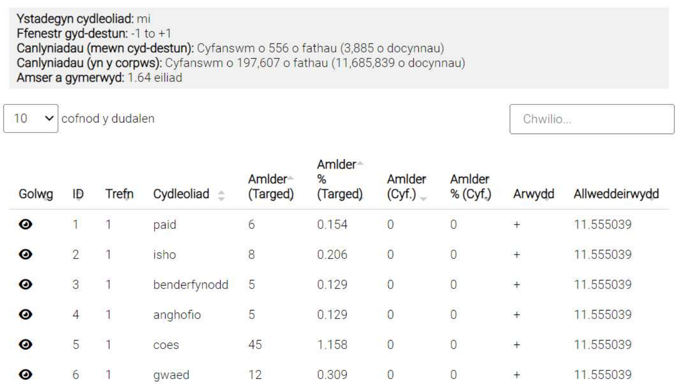
1. Dewis y tab **cydleoliad**. Yn y blwch **cydleoliadau**, teipiwch y gair rydych chi eisiau ei archwilio e.e. **tynnu**, cadwch y **ffenestr gyd-destun** i **+1 / -1** a chliciwch **dechrau**. I weld mwy o gofnodion, newidiwch **cofnod y dudalen** i **20.** Mae’r canlyniadau wedi’u trefnu yn ôl eu hallweddeirwydd, gan restru’r eitem fwyaf allweddol yn gyntaf (h.y. y cydleoliadau â’r berthynas gryfaf â **tynnu**).

2. Trafodwch y canlyniadau mewn grwpiau. Pam mae rhai o’r enghreifftiau mwyaf aml yn annisgwyl e.e. **yn, wedi, eu**? Ar wahân i’r rhain, allwch chi gytuno ar y pedwar cydleoliad mwyaf cyffredin i’ch gair chwilio? Yn achos **tynnu** byddai’r canlyniadau’n awgrymu:

- **tynnu sylw**
- **tynnu llun/iau**
- **tynnu coes**
- **tynnu allan**



|rhyngwyneb yr offeryn **cydleoliad**



Ystadegyn cydleoliad: mi
Ffenestr gyd-destun: -1 to +1
Canlyniadau (mewn cyd-destun): Cyfanswm o 556 o fathau (3,885 o docynnau)
Canlyniadau (yn y corpws): Cyfanswm o 197,607 o fathau (11,685,839 o docynnau)
Amser a gymerwyd: 1.64 eiliad

10 cofnod y dudalen — Chwilio...

Golwg	ID	Trefn	Cydleoliad	Amlder (Targed)	Amlder % (Targed)	Amlder (Cyf.)	Amlder % (Cyf.)	Arwydd	Allweddeirwydd
	1	1	paid	6	0.154	0	0	+	11.555039
	2	1	isho	8	0.206	0	0	+	11.555039
	3	1	benderfynodd	5	0.129	0	0	+	11.555039
	4	1	anghofio	5	0.129	0	0	+	11.555039
	5	1	coes	45	1.158	0	0	+	11.555039
	6	1	gwaed	12	0.309	0	0	+	11.555039

|canlyniadau cydleoliadau **tynnu**

3. Gyda’ch tiwtor ac aelodau eich grŵp, dewiswch bum gair rydych chi wedi eu dysgu ym Mynediad neu Sylfaen a dilynwch gamau 1 a 2 uchod i nodi rhai o’r prif gydleoliadau sy’n cael eu cysylltu â nhw.

Dysgwyr ac Athrawon

B2 Uwch: edrych ar ebychiadau (tasg enghreifftiol)

NOD: Mae'r dasg hon yn cyd-fynd â thrafod a gweithio ar ebychiadau (**ebych**). Mae'n annog myfyrwyr i ddarganfod mwy am (i) yr ebychiadau mwyaf cyffredin mewn Cymraeg llafar cyfoes (ii) ble maen nhw'n cael eu defnyddio a (iii) sut i'w hymgorffori yn eu repertoire ieithyddol eu hunain.

1. Agorwch offeryn **KWIC**, ac o dan **priodoleddau** dewiswch **Ebych**. Cliciwch y tab **hidlo metadata** ac o dan **modd**, cliciwch ar **llafar** i sicrhau mai enghreifftiau o ddata **llafar** yn unig a geir. Gellir dod o hyd i ebychiadau yn CorCenCC i gyd ond rydyn ni am ganolbwyntio ar sut maen nhw'n cael eu defnyddio yn yr iaith lafar. Cliciwch **dechrau**.

2. Trafodwch rai o'r ebychiadau mwyaf cyffredin rydych yn dod ar eu traws yma. Fe welwch fod rhai e.e. **ymm** yn nodi seibiau byr wrth siarad ac yn farcwyr sy'n cael eu defnyddio wrth drawsgrifio. Gwnewch restr o'r rhai mwyaf cyffredin a thrafodwch eu hystyron mewn grwpiau. Ble maen nhw'n fwyaf tebygol o ddigwydd – ar ddechrau, canol neu ddiwedd brawddeg? Oes ganddyn nhw swyddogaethau penodol mewn sgwrs e.e. nodi syndod neu sioc, cytuno neu oedi, ac os felly, pa rai?

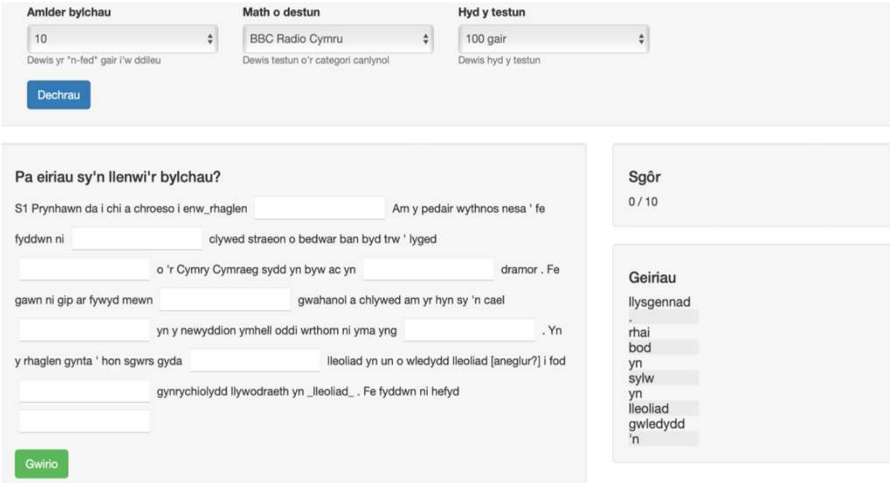
3. Dewiswch un neu ddau ebychiad i edrych arnyn nhw'n fwy manwl. Er enghraifft, gallen ni edrych ar **t'mod**. I wneud hyn, cadwch **ebych** yn y blwch **priodoleddau** gan ychwanegu **gwybod** yn y blwch **lema** a chlicio **dechrau**. Rhoddir pob ebychiad posibl wedi'i seilio ar **gwybod** yma. Faint o amrywiadau sydd? Trafodwch y rhain yn eich grŵp – oes unrhyw amrywiadau rhanbarthol / tafodieithol? Beth yw diben/ion y rhain?

4. Gallai rhan olaf y sesiwn ganolbwyntio ar sut gallai myfyrwyr ymgorffori ebychiadau yn eu disgwrs eu hunain. Faint/pa rai ydych chi'n eu defnyddio yn barod? Ydych chi wedi clywed siaradwyr Cymraeg eraill yn defnyddio'r ebychiadau hyn? Defnyddiwch rai o'r enghreifftiau rydych chi wedi dod ar eu traws yn **CorCenCC** yn sail i greu deialog fer (i) i ddangos rhai o'r ebychiadau rydych chi wedi'u hastudio ar waith a (ii) i esbonio pam maen nhw'n cael eu defnyddio.

Dysgwyr ac Athrawon

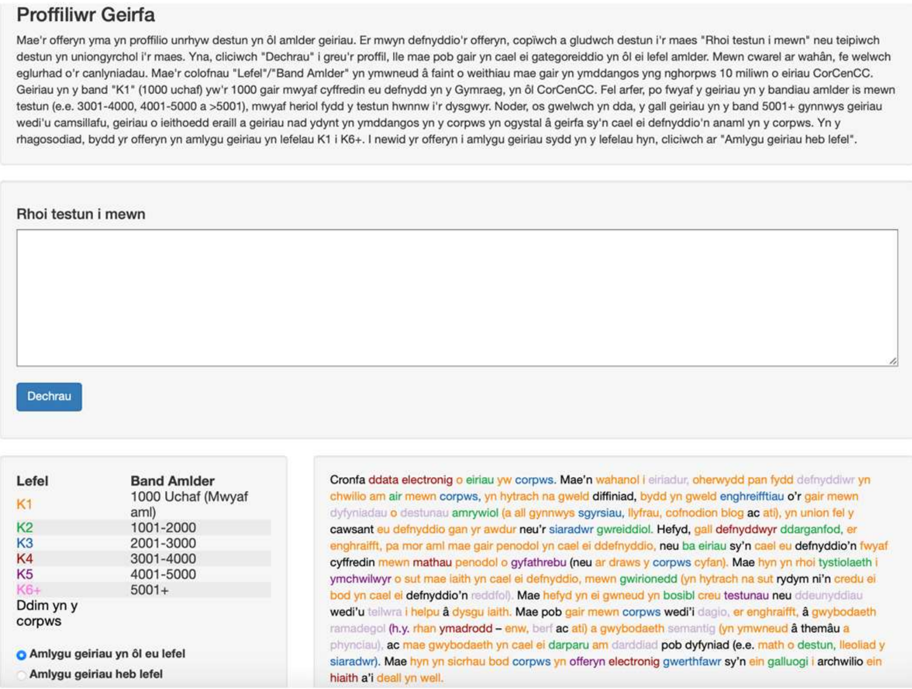
Y Tiwtiadur

Mae'r Tiwtiadur yn cynnwys offeryn **llenwi bylchau** lle mae geiriau'n cael eu dileu o destun ar gyfnodau penodol. Gall tiwtoriaid ddewis y cyfnodau, y math o destun a ddefnyddir a hyd y testun.



|**rhyngwyneb**
llenwi bylchau

Bydd angen i diwtoriaid addasu'r deunydd ar gyfer eu dosbarthiadau yn ogystal â newid neu ddisodli e.e. marcwyr anonymeiddio, fel **lleoliad** uchod.



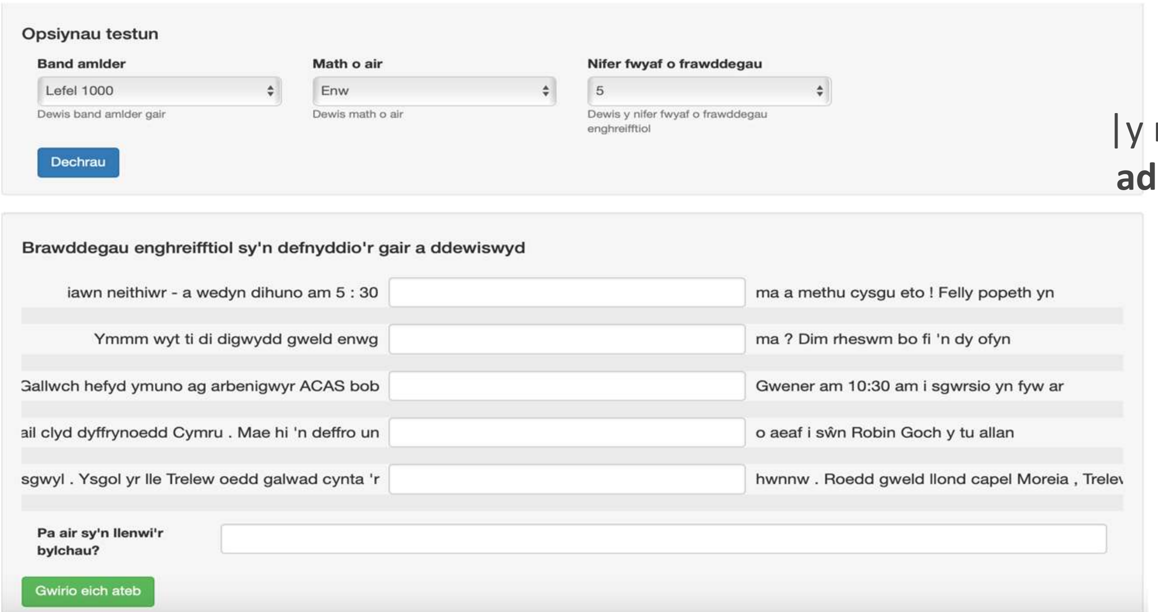
|rhyngwyneb **proffiliwr geirfa**

Hefyd, mae'r Tiwtiadur yn cynnwys offeryn **proffiliwr geirfa**. Mae hwn yn codio geiriau mewn testun mewn lliw yn ôl eu bandiau amlder. Mae hyn yn rhoi syniad i diwtoriaid o ba mor anodd neu hygyrch gall testun fod: po fwyaf y geiriau band amlder uchel mewn testun, hawsaf yn y byd ydy i ddarllenwyr eu deall.

Mae hyn hefyd yn debygol o fod o ddefnydd penodol i gyfieithwyr a chyhoeddwyr sydd am sicrhau eu bod yn paratoi cyfieithiadau neu ddeunydd darllen wedi'u hanelu at gynulleidfa eang. Gall unrhyw destun gael ei ddadansoddi'n gyflym trwy'r **proffiliwr geirfa.** Mae modd ystyried unrhyw eiriau amlder isel ac o bosibl eu haralleirio er mwyn gwneud yn siŵr fod y testun yn hawdd ei ddeall.

Dysgwyr ac Athrawon

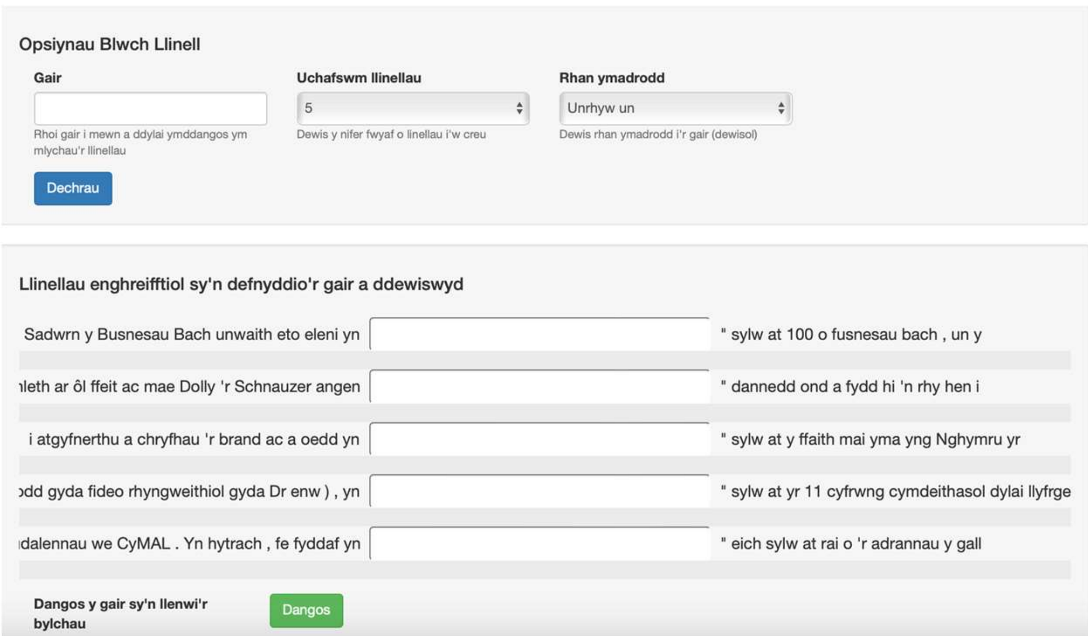
Mae’r Tiwtiadur yn cynnwys offeryn **adnabod geiriau** er mwyn dyfalu gair mewn cyd-destun. Gall tiwtoriaid amrywio’r band amlder, y math o air (Rhan Ymadrodd) a nifer fwyaf y geiriau yn yr offeryn. Mae dysgwyr yn teipio eu hateb ac yn ei wirio trwy glicio ar y botwm gwyrdd.



|y rhyngwyneb **adnabod geiriau**

Yn olaf, mae gan Y Tiwtiadur offeryn **creu tasg geiriau**. Gall tiwtoriaid guddio gair targed mewn llinellau concordans a gofynnir i ddysgwyr nodi beth sydd ar goll. Mae’r enghraifft isod yn ymestyn gwaith tasg cydleoliadau B1 ar dudalen 4 trwy dargedu’r gair **tynnu**. Gellir gosod uchafswm y llinellau a’r rhan ymadrodd er mwyn creu tasg bwrpasol.

|y rhyngwyneb **creu tasg geiriau**



Ymchwilwyr

Mae nifer o gymwysiadau ymchwil i CorCenCC. Er enghraifft:

Newidiadau gramadegol: fel rheol mae'r treiglad llaes yn digwydd ar ôl:

- y rhagenw meddiannol benywaidd **ei** ac unrhyw ffurfiau mewnol
- yr arddodiaid **â**, **gyda**, **tua**
- y cysyllteiriau **a, â**, **na, oni**
- y geirynnau negyddol **ni** a **na**

Er mwyn archwilio hyn, gwnewch chwiliad **KWIC** o'r lema **car** a dewis **llaes** o dan **treigladau**. Cliciwch **Hidlo Metadata,** dewiswch **modd**, a thrwy ddewis **llafar, ysgrifenedig** neu **electronig** a chwilio yn eu tro, gallwch weld enghreifftiau o'r treiglad llaes gyda'r lema **car** yn unol â phob modd. Wedyn, fe allech fod yn dymuno edrych ar y rhagenw meddiannol benywaidd **ei** ac a oes tystiolaeth i ddangos tuedd tuag at **car hi** (heb dreiglad llaes na rhagenw cychwynnol) neu **ei char hi** (yn cynnwys y treiglad llaes a'r rhagenw cychwynnol). Er mwyn gwneud hyn, yn y rhyngwyneb **uwch**, rhowch (i) **ei char** i mewn ac yna (ii) **car hi** (neu **ei car hi**). Gwelwn fod:

- **ei char** yn digwydd 15 gwaith yn CorCenCC
- **car hi** yn digwydd saith waith
- pob achos o **car hi** yn llafar
- y defnydd mwyaf o **ei char** yn ysgrifenedig (10) (2 yn electronig a 3 yn llafar)

Mae'r canlyniadau'n awgrymu bod tuedd i'r gystrawen heb dreiglad **car hi** fod yn fwy cynhyrchiol nag **ei char** ar lafar. Trwy hidlo'r metadata mae modd i ni edrych yn fwy manwl byth e.e. patrymau ar draws rhanbarthau penodol yng Nghymru.

Symud tuag at safoni geirfaol: nodir y rhain fel amrywiadau rhanbarthol cyffredin ar y gair Cymraeg am '*pancake*':

- **crempog(-ion)**
- **ffroesen/ffroes** neu **ffroisen/ffrois**
- **pancosen/pancos**
- **cramwythen/cramwyth**

Oes tystiolaeth yn CorCenCC i awgrymu bod symudiad tuag at un o'r rhain yn datblygu fel y gair 'safonol' am *pancake* yn Gymraeg?

Trwy wneud chwiliad **KWIC** am bob un o'r uchod, mae tystiolaeth glir fod **crempog** yn ymddangos fel y prif air o blith y rhain:

- 6 enghraifft o **ffroes** yn unig a geir, gyda phob un yn llafar (ac yn digwydd mewn trafodaethau am y gair Cymraeg am *pancake*!)
- mae **pancos** yn digwydd unwaith yn y data electronig, unwaith yn y data llafar a dwywaith yn y data ysgrifenedig
- nid oes cofnod o **cramwyth(en)**
- mae **crempog(-ion)** yn digwydd 93 o weithiau: 42 yn electronig, 39 yn llafar a 12 yn ysgrifenedig

Geiriadurwyr

Gan fod CorCenCC yn cynnwys data o ffynonellau llafar, electronig ac ysgrifenedig o Gymraeg Cyfoes, mae modd ystyried faint mae geiriau a fathwyd yn ddiweddar wedi lledaenu a dod o hyd i enghreifftiau o sut maen nhw'n cael eu defnyddio. Cymerwn y gair **hunlun**. Mae'r enghraifft gynharaf yn Geiriadur y Brifysgol yn dyddio o 2013. Trwy wneud chwiliadau **KWIC** a **chydleoliad** am y tri phosibilrwydd canlynol - **hunlun, selfie,** a **selffi** (does dim cofnod o **hun-lun**), mae'r canlyniadau'n dangos bod:

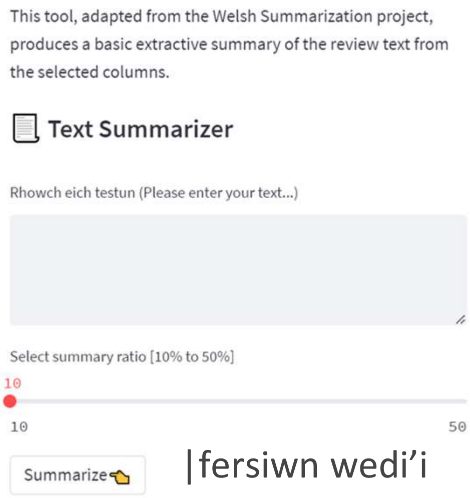
- **hunlun** yn digwydd yn y data ysgrifenedig ac electronig yn unig
- **selfie** ond yn digwydd yn y data electronig a llafar
- **selffi** ond yn cael ei ganfod yn y data electronig heb lawer o enghreifftiau
- cydleoliadau gyda **hunlun** yn awgrymu ei fod yn cael ei ddefnyddio'n debyg i **llun** gyda **tynnu hunlun**
- cydleoliadau gyda **selfie** yn awgrymu ei fod yn cael ei ddefnyddio yn y cyfuniad mwy seisnigedig **cymryd selfie**

Gair arall sydd wedi bod yn amlwg yn ddiweddar yw **brexit**. Mae dros 1,300 cofnod o **brexit** (wedi'i sillafu gydag 'x') a dim ond pump ar gyfer **bregsit**, gyda phob un yn codi o'r data llafar ac felly'n adlewyrchu orgraff y trawsgrifwyr. Mae'r dystiolaeth yn awgrymu, felly, fod yr opsiwn orgraffyddol Cymraeg arall o **bregsit** wedi'i ddiystyru o blaid y **brexit** gwreiddiol. Mae enghreifftiau fel **brexit caled** a **brexit di-gytundeb** yn dangos ei fod yn cael ei drin fel enw gwrywaidd yn y Gymraeg sy'n gwrthsefyll treiglad ar ddechrau'r gair (gw. **ar brexit / tuag at brexit**).

Cyfieithwyr

Gw. yr adran ar offeryn y **proffiliwr geirfa** ar dudalen 6. Unwaith i chi ddarllen yr eglurhad hwn am yr offeryn, gallech roi cynnig ar y **proffiliwr geirfa** gyda chyfieithiad wedi'i anelu at gynulleidfa gyhoeddus eang i asesu pa mor hygyrch ydyw yn ieithyddol. Er enghraifft, faint o eiriau amlder uchel ac amlder isel sy'n cael eu cynnwys yn eich cyfieithiad (neu gallwch gymharu â'r gwreiddiol os yw'n Saesneg trwy ddefnyddio **LexTutor:** https://www.lextutor.ca/vp/eng/).

Prosiectau Lloeren



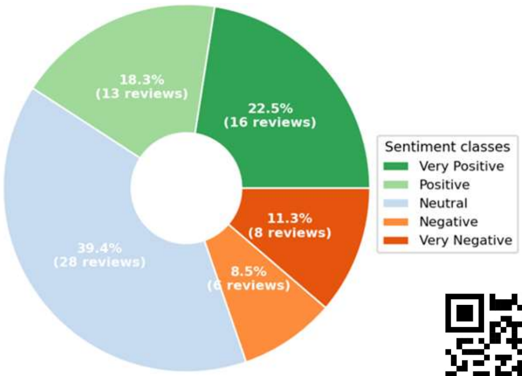
|fersiwn wedi'i haddasu o ACC (yn **TestunRhydd**)

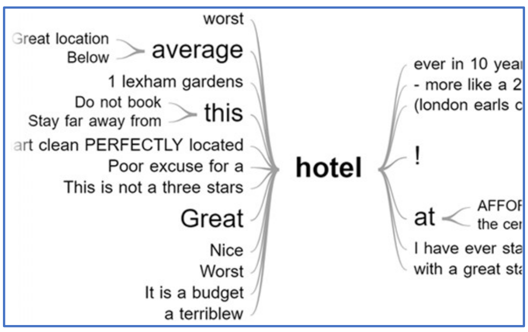
ACC: Crynhoi Testunau Cymraeg yn Awtomatig

Mae'r offeryn yma sydd ar gael yn rhad ac am ddim yn gadael i chi grynhoi dogfennau hir yn gyflym i'w cyflwyno'n effeithlon. Bydd **ACC** yn galluogi athrawon, er enghraifft, i addasu dogfennau hir i'w defnyddio yn yr ystafell ddosbarth, neu aelodau'r cyhoedd i greu a darllen crynodeb o wybodaeth gymhleth ar y rhyngrwyd. Ariannwyd prosiect ACC gan Lywodraeth Cymru. Am fwy o fanylion, ewch i wefan CorCenCC.

TestunRhydd

Mae **TestunRhydd** yn cefnogi dadansoddi a delweddu data testun rhydd yn y Gymraeg a'r Saesneg (e.e. data o arolygon, holiaduron, fforymau adborth). Mae TestunRhydd yn defnyddio rhai o'r gwasanaethau a'r methodolegau corpws o CorCenCC ac ACC (uchod), ac yn eu hailbecynnu fel bod modd i gynulleidfaoedd a grwpiau o ddefnyddwyr newydd ddadansoddi eu data adborth eu hunain. Cafodd ei gynllunio ar y cyd ag Ymddiriedolaeth Genedlaethol Cymru, Amgueddfa Cymru, Cadw, CBAC a'r Ganolfan Dysgu Cymraeg Genedlaethol. Cyllidwyd Testun Rhydd gan AHRC, ac mae ar gael i *unrhyw un* mewn *unrhyw sector* yng Nghymru a'r tu hwnt.





|delweddau sampl yn **TestunRhydd**

Thesawrws

Mae tîm prosiect **Thesawrws** a ariennir gan Lywodraeth Cymru yn datblygu thesawrws ar-lein o'r Gymraeg (wedi'i adeiladu'n rhannol gan ddefnyddio set ddata CorCenCC) sydd ar gael am ddim ac yn fodd i chi chwilio am gyfystyron (h.y. dau neu fwy o eiriau â bron yr un ystyr). Am fwy o fanylion, ewch i wefan CorCenCC.

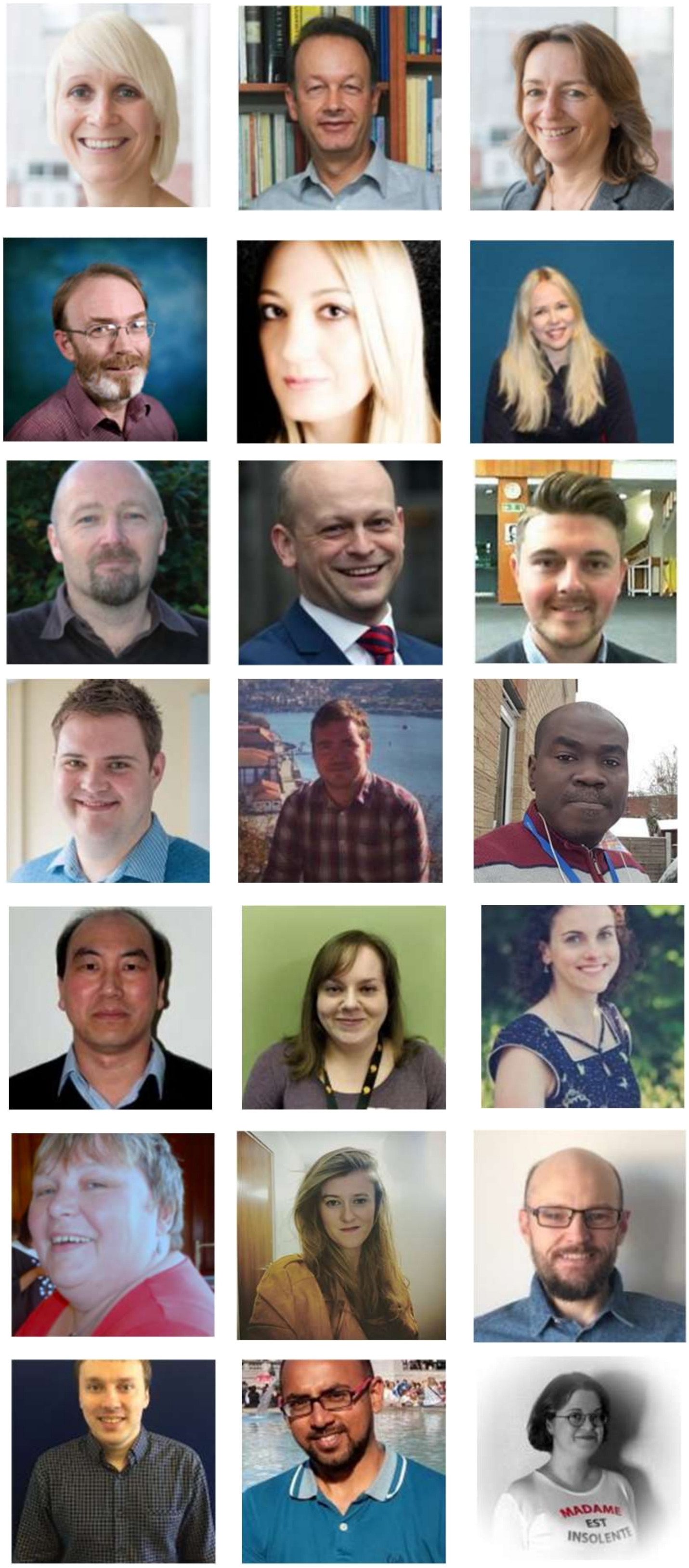
Darganfod Mwy

Am fwy o fanylion am CorCenCC, cyhoeddiadau sy'n gysylltiedig â'r prosiect a phrosiectau lloeren, ewch i'r brif wefan yn www.corcencc.cymru

Mae canllawiau fideo ar gael yn www.youtube.com/@corcencc

Hefyd gallwch ein dilyn ar Twitter @CorCenCC

I gysylltu â'r tîm, ebostiwch, CorCenCC@caerdydd.ac.uk



Roedd CorCenCC yn brosiect ymchwil a gyllidwyd gan ESRC/AHRC (Rhif y Grant ES/M011348/1). Roedd y tîm yn cynnwys **PY** - Dawn Knight; **CYwyr -** Tess Fitzpatrick, Steve Morris, Irena Spasić, Paul Rayson, Enlli Thomas, Mark Stonelake, Alex Lovell, Jonathan Morris and Jeremy Evas; **CY** - Steven Neale, Jennifer Needs, Mair Rees, Scott Piao, Ignatius Ezeani a Gareth Watkins; **Myfyrwyr PhD** - Vigneshwaran Muralidaran a Bethan Tovey-Walsh; **Ymgynghorwyr** - Kevin Donnelly, Kevin Scannell, Laurence Anthony, Tom Cobb, Michael McCarthy a Margaret Deuchar; **Grŵp Ymgynghorol y Prosiect** - Colin Williams, Karen Corrigan, Llion Jones, Maggie Tallerman, Mair Parry-Jones, Gwen Awbery, Emyr Davies, Gareth Morlais, Owain Roberts, Aran Jones ac Andrew Hawke.