Economic and Social Research Council

Arts and Humanities Research Council

www.corcencc.org

CorCenCC

Corpws Cenedlaethol Cymraeg Cyfoes

National Corpus of Contemporary Welsh

Cardiff University
Prifysgol Caerdydd

Swansea University
Prifysgol Abertawe

Prifysgol Bangor University

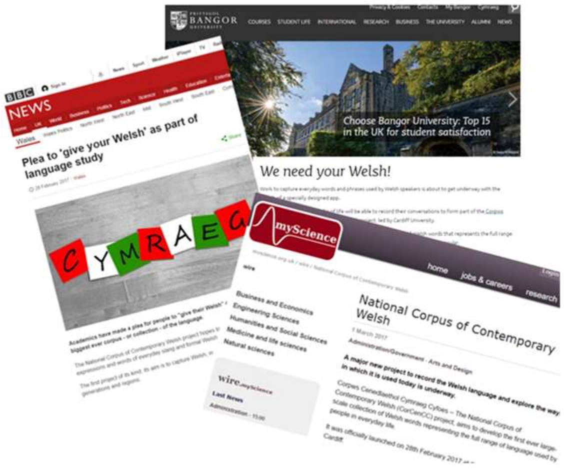
Lancaster University

Overview

CorCenCC is a freely accessible collection of samples of Welsh, gathered from real-life communication, and stored as an electronic database (a 'corpus'). You can search CorCenCC to find out about how people really use Welsh, for instance, how frequently a specific word is used, or what are the most frequently used words in specific kinds of communication (or across the entire corpus).

CorCenCC contains over 11 million words from written, spoken and electronic (online, digital texts) Welsh language sources, taken from a range of genres, language varieties (regional and social) and contexts (see page 2 for more details). Every word in a corpus is 'tagged' with, for example, grammatical information (i.e. part of speech – noun, verb, etc.) and semantic information (relating to themes and topics), and information is provided about where each language excerpt is from (e.g. text type, speaker location). This makes CorCenCC a valuable electronic tool for allowing us to explore and to better understand our language.

This booklet demonstrates who might benefit from using CorCenCC and how the corpus can be used, as well as briefly profiling some recent satellite projects that have emerged following the development of this national resource.



Plea to 'give your Welsh' as part of language study

28 February 2017 Wales

Academics have made a plea for people to "give their Welsh"
biggest ever corpus - or collection - of the language.

The National Corpus of Contemporary Welsh project hopes to
expressions and words of everyday slang and formal Welsh.

The first project of its kind, its aim is to capture Welsh, in
generations and regions.

Choose Bangor University: Top 15 in the UK for student satisfaction

We need your Welsh!

Work to capture everyday words and phrases used by Welsh speakers is about to get underway with the
a specially designed app.

National Corpus of Contemporary Welsh

1 March 2017

Administration/Government - Arts and Design

A major new project to record the Welsh language and explore the way in which it is used today is underway.

Corpws Cenedlaethol Cymraeg Cyfoes – The National Corpus of Contemporary Welsh (CorCenCC) project, aims to develop the first ever large-scale collection of Welsh words representing the full range of language used by people in everyday life.

It was officially launched on 28th February 2017 at
Cardiff.

Contents

|Page 1: Overview

|Page 2: Data Details

|Page 3: Features: v2.0

|Pages 4-7:
Learners & Teachers

|Page 8: Researchers

|Page 9:
Lexicographers & Translators

|Page 10:
Satellite Projects

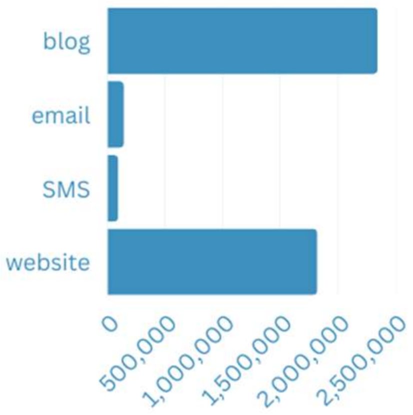
|Page 11: Find Out More

Data Details

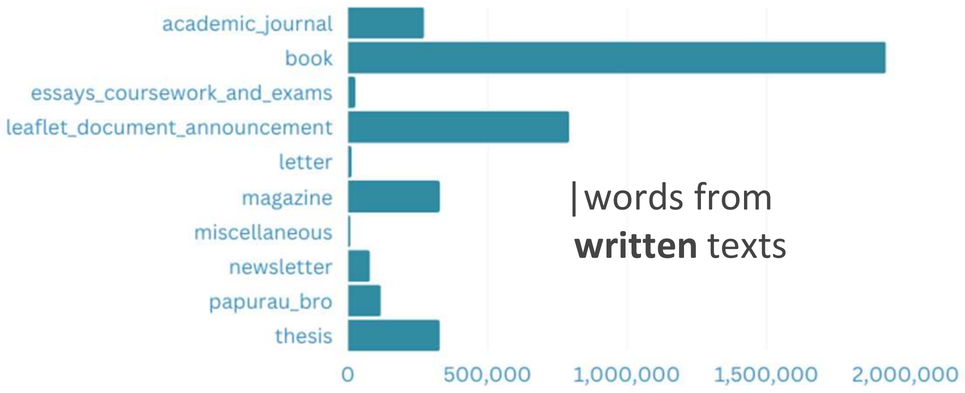
|3.9m words from **written** texts

|4.4m words from **e-language** (digitally created texts)

|2.86m words from **spoken** language



|word counts of **e-language** texts

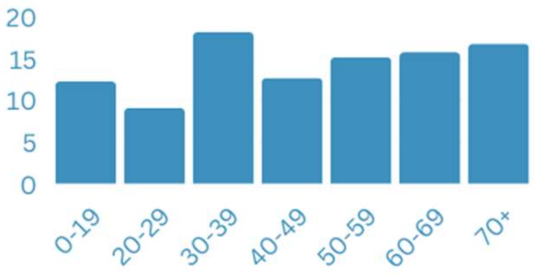


|words from **written** texts

|of all **written** books:

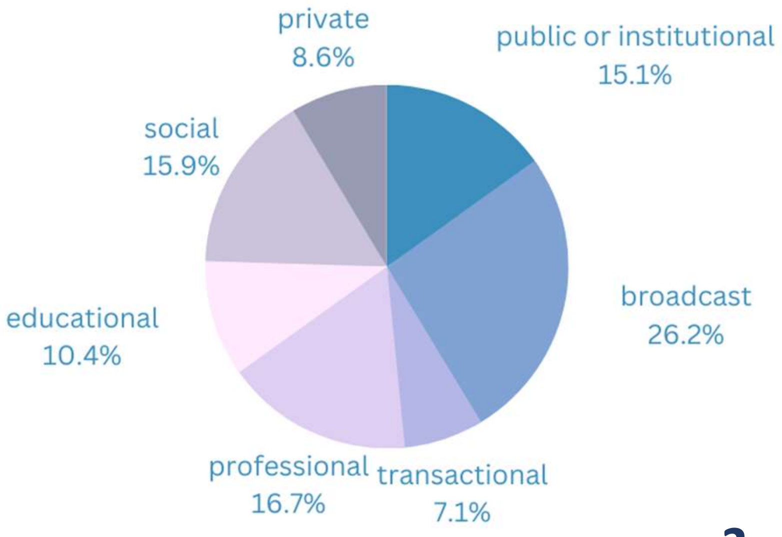
> 22% are for adults
> 11.25% are for children
> 4% are for adult learners
> 3.5% are for young learners

|ages of **spoken** data contributors (by %)



|of contributors of **spoken** data identified as female

|recording contexts of **spoken** data

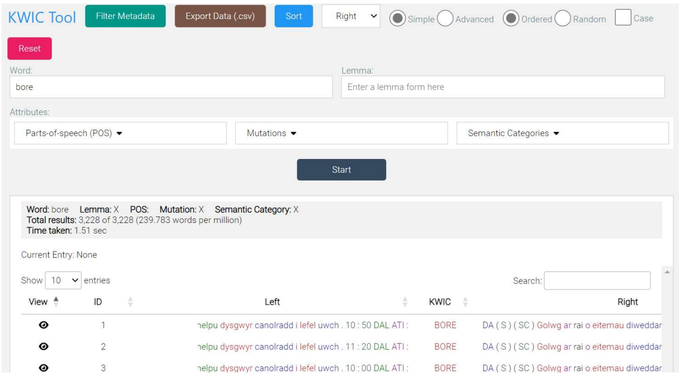


Infographics produced in Canva.com

Features: v2.0

KWIC (Key Word In Context)

KWIC is the default way in to exploring CorCenCC. You can type any word, phrase, grammatical (part-of-speech - POS) or semantic category into the search box in this tool and find out how often it occurs either in the entire corpus, or in a specific subset of corpus data. You can filter by restricting your search to different modes (written, spoken, e-language), genre(s), context(s), by contributors' age, gender, or location, for example. You will see examples of the word you searched for in context, as it is used, in the KWIC display. The KWIC tool also tells you how often your search term is used in the dataset.



Word: bore

Lemma: Enter a lemma form here

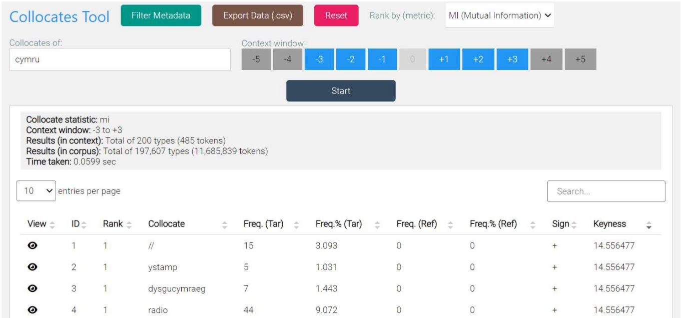
Word: bore **Lemma:** X **POS:** **Mutation:** X **Semantic Category:** X
Total results: 3,228 of 3,228 (239.783 words per million)
Time taken: 1.51 sec

Current Entry: None

Show 10 entries

View	ID	Left	KWIC	Right
👁	1	ıelpu dysgwyr canolradd i lefel uwch . 10 : 50 DAL ATI :	BORE	DA (S) (SC) Golwg ar rai o eitemau diweddar
👁	2	ıelpu dysgwyr canolradd i lefel uwch . 11 : 20 DAL ATI :	BORE	DA (S) (SC) Golwg ar rai o eitemau diweddar
👁	3	ıelpu dysgwyr canolradd i lefel uwch . 10 : 00 DAL ATI :	BORE	DA (S) (SC) Golwg ar rai o eitemau diweddar

|**KWIC** (above) & **collocate** (below) tools



Rank by (metric): MI (Mutual Information)

Collocates of: cymru

Context window: -5 -4 -3 -2 -1 0 +1 +2 +3 +4 +5

Collocate statistic: mi
Context window: -3 to +3
Results (in context): Total of 200 types (485 tokens)
Results (in corpus): Total of 197,607 types (11,685,839 tokens)
Time taken: 0.0599 sec

10 entries per page

View	ID	Rank	Collocate	Freq. (Tar)	Freq.% (Tar)	Freq. (Ref)	Freq.% (Ref)	Sign	Keyness
👁	1	1	//	15	3.093	0	0	+	14.556477
👁	2	1	ystamp	5	1.031	0	0	+	14.556477
👁	3	1	dysgucymraeg	7	1.443	0	0	+	14.556477
👁	4	1	radio	44	9.072	0	0	+	14.556477

Frequency List

You can find out how often different words and/or lemmas (i.e. the base of a word and its inflectional forms) occur in the entire CorCenCC dataset, or in subsets of the data (see **KWIC**). Frequency lists of grammatical categories (POS), mutation types and semantic categories can also be created.

N-grams

As with the **frequency** list tool, you can generate lists of common clusters (n-grams) of words that occur in the corpus, from the most common two words that co-occur, to patterns of five words.

Collocates

Collocation refers to the relationship between words and the relative attraction between them. You can search for **collocates** of a given word (or POS/semantic tag) within a specific window, i.e. up to 5 words on either side of that search word.

Keyword

Keyword analysis allows you to compare different subsets of the corpus data against each other to determine which words, phrases, tags etc. occur significantly more often in one subset compared to the other.

Learners & Teachers

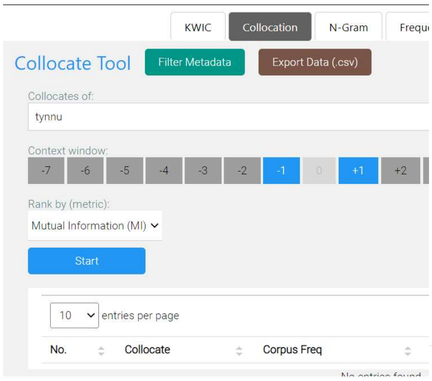
B1 Canolradd: collocations (example task)

AIM: By using the **collocation** analysis tool, students will learn how to identify common collocations associated with different search words that they are interested in (in this case, **tynnu**). This is a way to expand students' lexical knowledge based on words they have learnt at A1 and A2 levels.

1. Select the **collocates** tab. In the **collocates of** box, type the word you want to investigate e.g. **tynnu** [*to pull*], keep the **context window** and click **start**. To view more entries, change the **entries per page** to 20. The results are ranked according to their keyness, listing the most key item first (i.e. the collocates with the strongest relationship to **tynnu**).

2. Discuss the results in groups. Why are some of the most frequent examples unexpected e.g. **yn, wedi, eu**? Apart from these, can you agree on the four most common collocations for your search word? In the case of **tynnu** the results would suggest:

- **tynnu sylw** [*to draw attention*]
- **tynnu llun/iau** [*to take a picture*]
- **tynnu coes** [*to pull someone's leg*]
- **tynnu allan** [*to pull out*]



Collocates of: tynnu

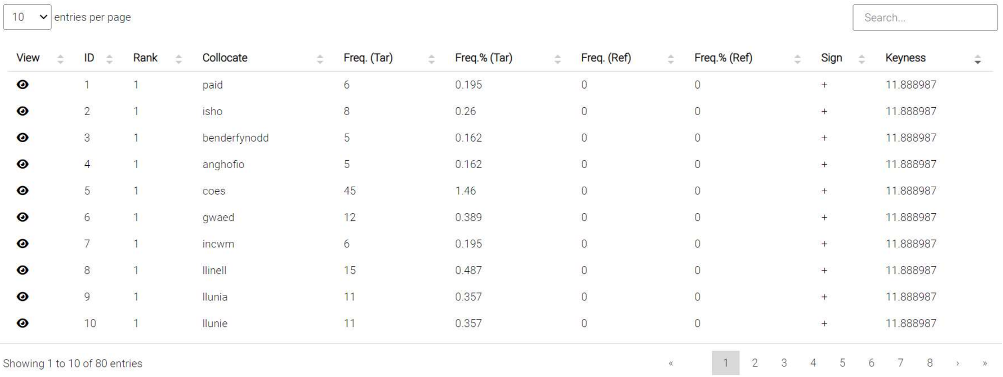
Context window: -7 -6 -5 -4 -3 -2 -1 0 +1 +2

Rank by (metric): Mutual Information (MI)

10 entries per page

No entries found

|**collocate** tool interface



10 entries per page

View	ID	Rank	Collocate	Freq. (Tar)	Freq.% (Tar)	Freq. (Ref)	Freq.% (Ref)	Sign	Keyness
👁	1	1	paid	6	0.195	0	0	+	11.888987
👁	2	1	isho	8	0.26	0	0	+	11.888987
👁	3	1	benderfynodd	5	0.162	0	0	+	11.888987
👁	4	1	anghofio	5	0.162	0	0	+	11.888987
👁	5	1	coes	45	1.46	0	0	+	11.888987
👁	6	1	gwaed	12	0.389	0	0	+	11.888987
👁	7	1	incwm	6	0.195	0	0	+	11.888987
👁	8	1	llinell	15	0.487	0	0	+	11.888987
👁	9	1	llunia	11	0.357	0	0	+	11.888987
👁	10	1	llunie	11	0.357	0	0	+	11.888987

Showing 1 to 10 of 80 entries

« 1 2 3 4 5 6 7 8 › »

|collocate search results for **tynnu**

3. With your tutor and members of your group, select five words that you learnt in Mynediad or Sylfaen and follow steps 1 and 2 above to identify some of the main collocations associated with them.

Learners & Teachers

B2 Uwch: looking at interjections (example task)

AIM: This task complements discussion and work around interjections (**ebych**) in Welsh. It encourages students to find out more about (i) the most common interjections in contemporary spoken Welsh (ii) where they are used and (iii) how to incorporate them into their own linguistic repertoire.

1. Open the **KWIC** tool, and under **attributes** select **Ebych**. Click the **filter metadata** tab and under **mode**, click on **spoken** to ensure that all the examples are from the **spoken** data only. Interjections can be found in all of CorCenCC but we want to focus on how they are used in spoken language. Click **start**.

2. Discuss some of the most common interjections you observe here. You will find that some e.g. **ymm** indicate short pauses in speech and are markers used in transcription. Make a list of the most common ones and discuss their meanings in groups. Where are they most likely to occur - at the beginning, middle or the end of a sentence? Do any of these interjections perform specific roles in conversation e.g. surprise or shock, agreement or hesitation and if so, which?

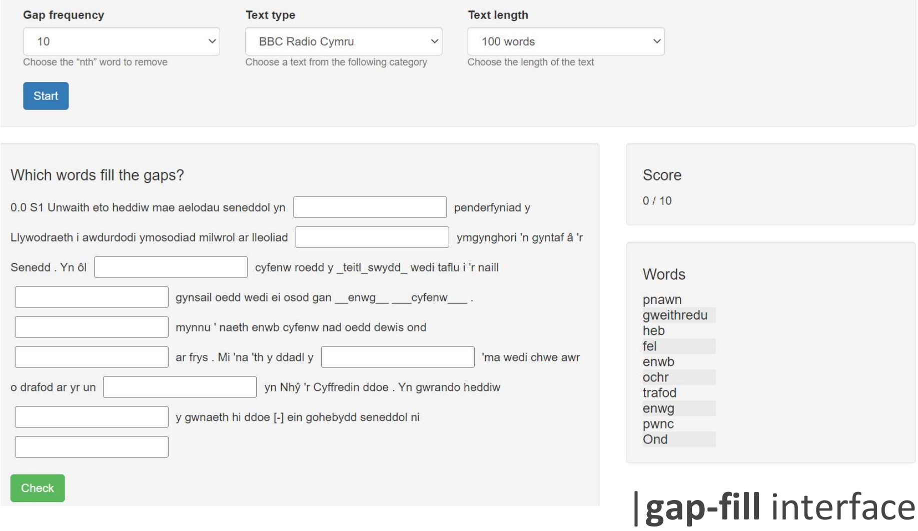
3. Choose one or two interjections to examine in more detail. For example, we could look at **t'mod** [*You know*]. To do this, keep **ebych** in the **attributes** box but add **gwybod** in the lemma box and click **start**. All possible interjections based on **gwybod** are given here. How many variations are there? Discuss these in your group - are there any regional/dialect variations? What purpose/s do these serve?

4. The final part of the session might focus on how students might incorporate interjections in their own discourse. How many/which ones do you already use? Have you observed other speakers of Welsh using these interjections? Use some of the examples you have found in **CorCenCC** as the basis for constructing a short dialogue (i) to show some of the interjections you have studied in use and (ii) to explain why they are used.

Learners & Teachers

Y Tiwtiadur

Y Tiwtiadur includes a **gap-fill** or cloze exercise creator, where words are deleted from a text at specified intervals. Tutors can select gap intervals, the type of text used and the length of the text.



Gap frequency
10
Choose the "nth" word to remove

Text type
BBC Radio Cymru
Choose a text from the following category

Text length
100 words
Choose the length of the text

Which words fill the gaps?

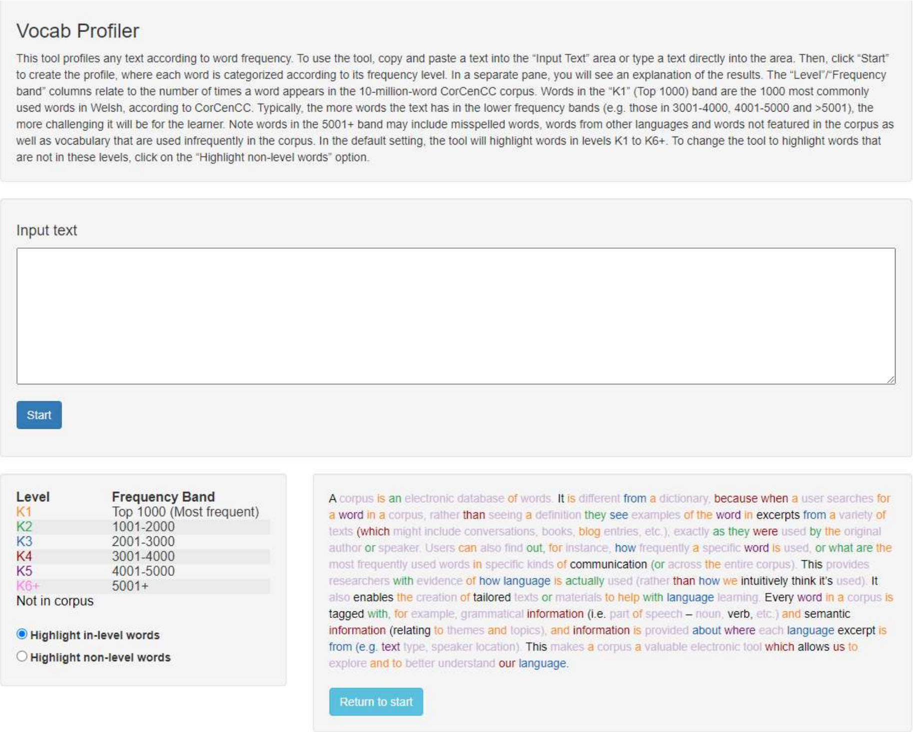
0.0 S1 Unwaith eto heddiw mae aelodau seneddol yn penderfyniad y
Llywodraeth i awdurdodi ymosodiad milwrol ar lleoliad ymgynghori 'n gyntaf â 'r
Senedd . Yn ôl cyfenw roedd y _teitl_swydd_ wedi taflu i 'r naill
gynsail oedd wedi ei osod gan __enwg__ ___cyfenw___ .
mynnu ' naeth enwb cyfenw nad oedd dewis ond
ar frys . Mi 'na 'th y ddadl y 'ma wedi chwe awr
o drafod ar yr un yn Nhŷ 'r Cyffredin ddoe . Yn gwrando heddiw
y gwnaeth hi ddoe [-] ein gohebydd seneddol ni

Score
0 / 10

Words
pnawn
gweithredu
heb
fel
enwb
ochr
trafod
enwg
pwnc
Ond

|**gap-fill** interface

Tutors will need to adapt the materials for their classes in addition to changing or substituting, e.g. anonymisation markers such as **lleoliad** [*location*] above.



Vocab Profiler

This tool profiles any text according to word frequency. To use the tool, copy and paste a text into the "Input Text" area or type a text directly into the area. Then, click "Start" to create the profile, where each word is categorized according to its frequency level. In a separate pane, you will see an explanation of the results. The "Level"/"Frequency band" columns relate to the number of times a word appears in the 10-million-word CorCenCC corpus. Words in the "K1" (Top 1000) band are the 1000 most commonly used words in Welsh, according to CorCenCC. Typically, the more words the text has in the lower frequency bands (e.g. those in 3001-4000, 4001-5000 and >5001), the more challenging it will be for the learner. Note words in the 5001+ band may include misspelled words, words from other languages and words not featured in the corpus as well as vocabulary that are used infrequently in the corpus. In the default setting, the tool will highlight words in levels K1 to K6+. To change the tool to highlight words that are not in these levels, click on the "Highlight non-level words" option.

Input text

Level	Frequency Band
K1	Top 1000 (Most frequent)
K2	1001-2000
K3	2001-3000
K4	3001-4000
K5	4001-5000
K6+	5001+
Not in corpus	

● Highlight in-level words
○ Highlight non-level words

A corpus is an electronic database of words. It is different from a dictionary, because when a user searches for a word in a corpus, rather than seeing a definition they see examples of the word in excerpts from a variety of texts (which might include conversations, books, blog entries, etc.), exactly as they were used by the original author or speaker. Users can also find out, for instance, how frequently a specific word is used, or what are the most frequently used words in specific kinds of communication (or across the entire corpus). This provides researchers with evidence of how language is actually used (rather than how we intuitively think it's used). It also enables the creation of tailored texts or materials to help with language learning. Every word in a corpus is tagged with, for example, grammatical information (i.e. part of speech – noun, verb, etc.) and semantic information (relating to themes and topics), and information is provided about where each language excerpt is from (e.g. text type, speaker location). This makes a corpus a valuable electronic tool which allows us to explore and to better understand our language.

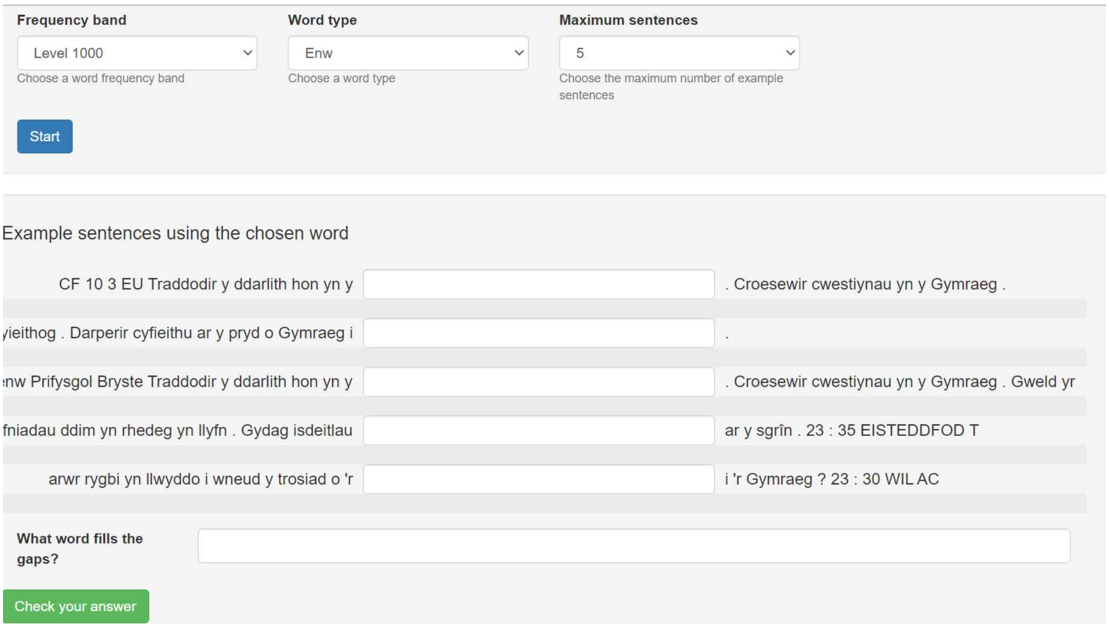
|**vocab profiler** interface

Y Tiwtiadur also includes a **vocab profiler** tool. This tool colour codes words in a text according to their different frequency bands. This gives tutors an idea of how difficult or accessible a text might be: the more higher frequency band words contained within a text, the easier it will be for readers to understand them.

This is also likely to be of particular use for translators and publishers who are concerned with preparing translations or reading material aimed at a wide audience. Any text can be quickly analysed through the **vocab profiler** and any low frequency words considered and possibly paraphrased to maximise comprehension.

Learners & Teachers

Y Tiwtiadur also includes a **word identifier** tool for guessing a word in context. Tutors can vary the frequency band, the kind of word (POS) and the maximum number of words provided by the tool. Learners type in their answer and check it by clicking on the green button.



Frequency band
Level 1000
Choose a word frequency band

Word type
Enw
Choose a word type

Maximum sentences
5
Choose the maximum number of example sentences

Start

Example sentences using the chosen word

CF 10 3 EU Traddodir y ddarlith hon yn y ________ . Croesewir cwestiynau yn y Gymraeg .

ıieithog . Darperir cyfieithu ar y pryd o Gymraeg i ________ .

ınw Prifysgol Bryste Traddodir y ddarlith hon yn y ________ . Croesewir cwestiynau yn y Gymraeg . Gweld yr

fniadau ddim yn rhedeg yn llyfn . Gydag isdeitlau ________ ar y sgrîn . 23 : 35 EISTEDDFOD T

arwr rygbi yn llwyddo i wneud y trosiad o 'r ________ i 'r Gymraeg ? 23 : 30 WIL AC

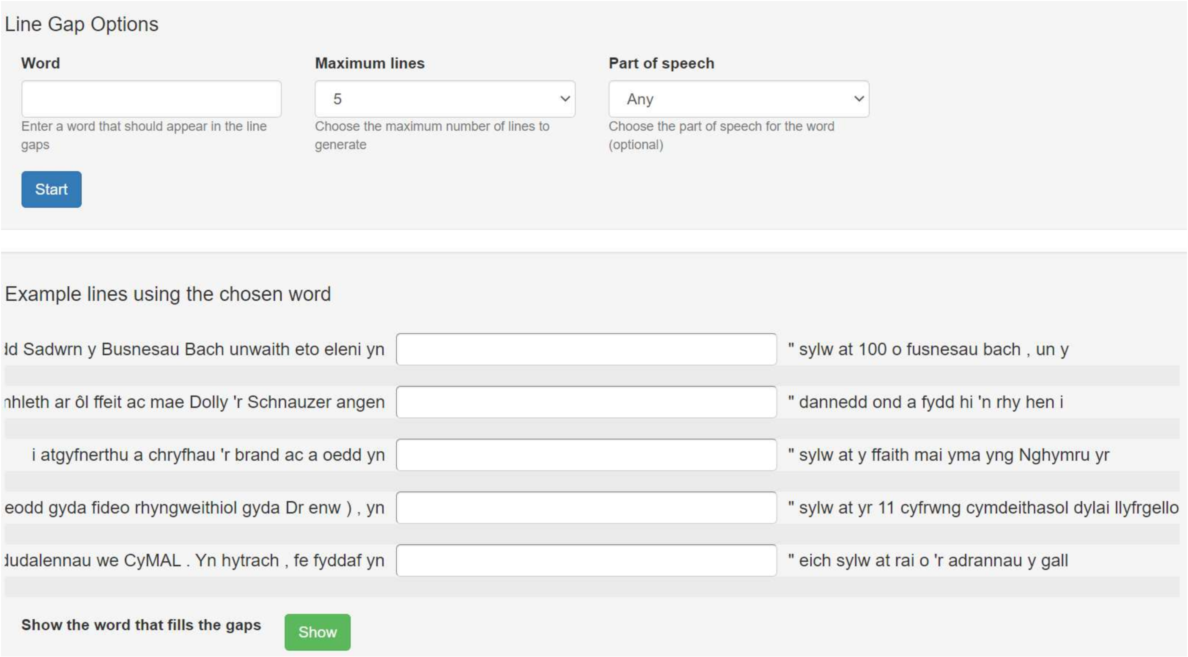
What word fills the gaps?

Check your answer

|**word identifier** interface

Finally, Y Tiwtiadur contains a **word task creator** tool. Tutors can blank out a target word in concordance lines and learners have to identify what is missing. The example below expands the work of the B1 collocation task on page 4 targeting the word **tynnu**. Both the number of lines and the POS can be set to create a bespoke task.

|**word task creator** interface



Line Gap Options

Word
Enter a word that should appear in the line gaps

Maximum lines
5
Choose the maximum number of lines to generate

Part of speech
Any
Choose the part of speech for the word (optional)

Start

Example lines using the chosen word

ıd Sadwrn y Busnesau Bach unwaith eto eleni yn ________ " sylw at 100 o fusnesau bach , un y

ınhleth ar ôl ffeit ac mae Dolly 'r Schnauzer angen ________ " dannedd ond a fydd hi 'n rhy hen i

i atgyfnerthu a chryfhau 'r brand ac a oedd yn ________ " sylw at y ffaith mai yma yng Nghymru yr

eodd gyda fideo rhyngweithiol gyda Dr enw) , yn ________ " sylw at yr 11 cyfrwng cymdeithasol dylai llyfrgello

ıudalennau we CyMAL . Yn hytrach , fe fyddaf yn ________ " eich sylw at rai o 'r adrannau y gall

Show the word that fills the gaps Show

Researchers

There are numerous research applications for CorCenCC. For example:

Grammatical changes: aspirate mutations typically occur after:

- feminine possessive **ei** [*her*] and any infixed forms
- prepositions **â** [*with*], **gyda** [*with*], **tua** [*towards*]
- conjunctions **a** [*and*], **â** [*as*], **na** [*than*], **oni** [*unless*]
- negative particles **ni** and **na**

To explore this, do a **KWIC** search of the lemma **car** [*car*] and under **mutations**, select **aspirate**. Click Filter Metadata, select mode, and by choosing spoken, written or electronic and searching in turn, you can view examples of aspirate mutations with the lemma **car** according to each mode. You might then look at the feminine possessive **ei** and whether there is evidence of a tendency towards **car hi** (no aspirate mutation, no initial pronoun) or **ei char hi** (aspirate mutation, initial pronoun included). To do this, in the **advanced** interface, input (i) **ei char** then (ii) **car hi** (or **ei car hi**).

This tells us that:

- **ei char** occurs 15 times in CorCenCC
- **car hi** occurs seven times
- all uses of **car hi** are spoken
- most use of **ei char** are written (10) (two are electronic, three spoken)

The results suggest there is tendency for the non-mutated **car hi** construction to be more productive than **ei char** in spoken Welsh. Filtering the metadata allows us to drill down into this even further e.g. usage patterns in particular regions of Wales.

Shifts towards lexical standardisation: regional variations for the word pancake in Welsh are commonly identified as:

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

Is there any evidence in CorCenCC to suggest that there is a shift towards one of these becoming the 'standard' word for *pancake* in Welsh?

By doing a **KWIC** search of each of the above, there is clear evidence that **crempog** is emerging as the principal Welsh equivalent of *pancake*:

- there are only six examples of **ffroes**, all of which are spoken (and occur in discussions about the Welsh word for *pancake*!)
- **pancos** occurs once in the electronic data, once in the spoken data and twice in the written data
- **cramwyth(en)** does not occur at all
- **crempog(-ion)** occurs 93 times: 42 electronic, 39 spoken and 12 written

Lexicographers

As CorCenCC contains data from spoken, electronic and written sources of contemporary Welsh, it is possible to consider the spread of recently coined words and find examples of how they are being used. Let's take the word **hunlun** [*selfie*]. The earliest example for this in the Dictionary of the Welsh Language dates from 2013. By doing **KWIC** and **collocation** searches for the following three possibilities for '*selfie'* in Welsh - **hunlun, selfie,** and **selffi** (there are no records of **hun-lun**), results show:

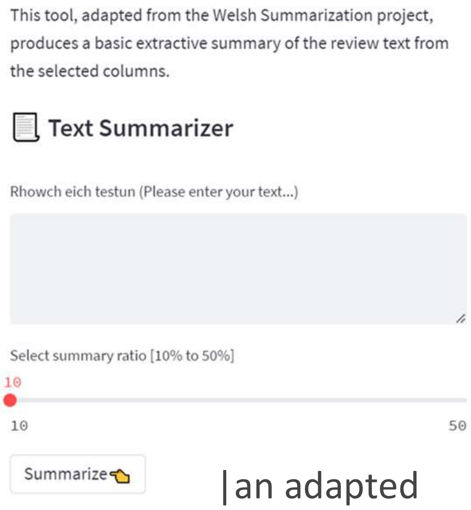
- **hunlun** occurs in written and electronic data only
- **selfie** only occurs in electronic and spoken data
- **selffi** is only found in electronic data with very few examples
- collocations with **hunlun** suggest that it is being used similarly to **llun** [*picture*] with **tynnu hunlun** [*to take - lit. to pull - a selfie*]
- collocations with **selfie** suggest that it is being used in the more anglicised **cymryd selfie** [*to take a selfie*]

Another word which has featured frequently recently is **brexit**. There are over 1,300 entries with (the spelling) **brexit** but only five for **bregsit**, all of which are from spoken data and therefore reflect the orthography of the transcribers. The evidence suggests, therefore, that the possible Welsh orthographical alternative **bregsit** has been dismissed for the **brexit** original. Examples such as **brexit caled** [*hard brexit*] and **brexit di-gytundeb** [*no-agreement brexit*] indicate that it is treated as a masculine noun in Welsh which resists initial mutation (cf **ar brexit / tuag at brexit**).

Translators

See the section on the **vocab profiler** tool on page 6. Once you have read through this explanation of the tool, you could try the **vocab profiler** out with a translation aimed at a wide public audience to assess how linguistically accessible it is. For example, how many high frequency and low frequency words are contained within your translation (or compare with the original if it is in English, using **LexTutor:** https://www.lextutor.ca/vp/eng/).

Satellite Projects



This tool, adapted from the Welsh Summarization project, produces a basic extractive summary of the review text from the selected columns.

Text Summarizer
Rhowch eich testun (Please enter your text...)
Select summary ratio [10% to 50%]
10
10
50
Summarize

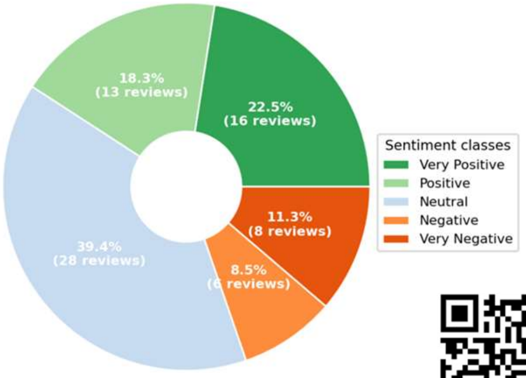
|an adapted version of ACC (in **FreeTxt**)

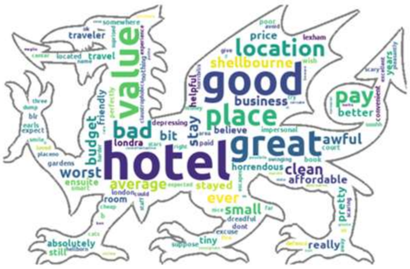
ACC: Welsh Automatic Text Summarisation

This freely available tool allows you to quickly summarise long documents for efficient presentation. **ACC** allows teachers, for example, to adapt long documents for use in the classroom, or members of the public to create and read a summary of complex information presented on the internet. The ACC project was funded by the Welsh Government. For more details, visit the CorCenCC website.

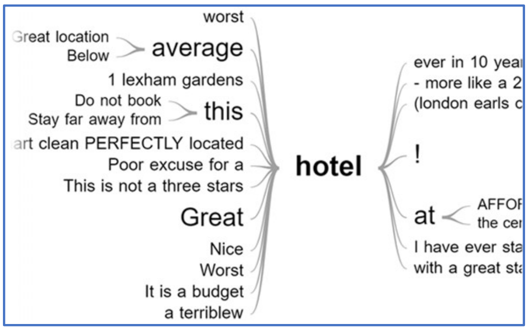
FreeTxt

FreeTxt supports the analysis and visualisation of free-text data in both English and Welsh (e.g. data from surveys, questionnaires, feedback fora). FreeTxt draws on some of the corpus-based utilities and methodologies from CorCenCC and ACC (above), repackaging these to enable new audiences and user-groups to analyse their own feedback data. Co-designed in collaboration with National Trust Wales, National Museum Wales, Cadw, WJEC, and National Centre for Learning Welsh, and funded by AHRC, FreeTxt is accessible to *anyone* in *any sector* in Wales and beyond.





|sample visualisations in **FreeTxt**



Thesawrws

The Welsh Government funded **Thesawrws** project team is developing a freely available online thesaurus of the Welsh language (built, in part, using the CorCenCC dataset), enabling you to search for synonyms (i.e. two or more words with nearly the same meaning). For more details, visit the CorCenCC website.

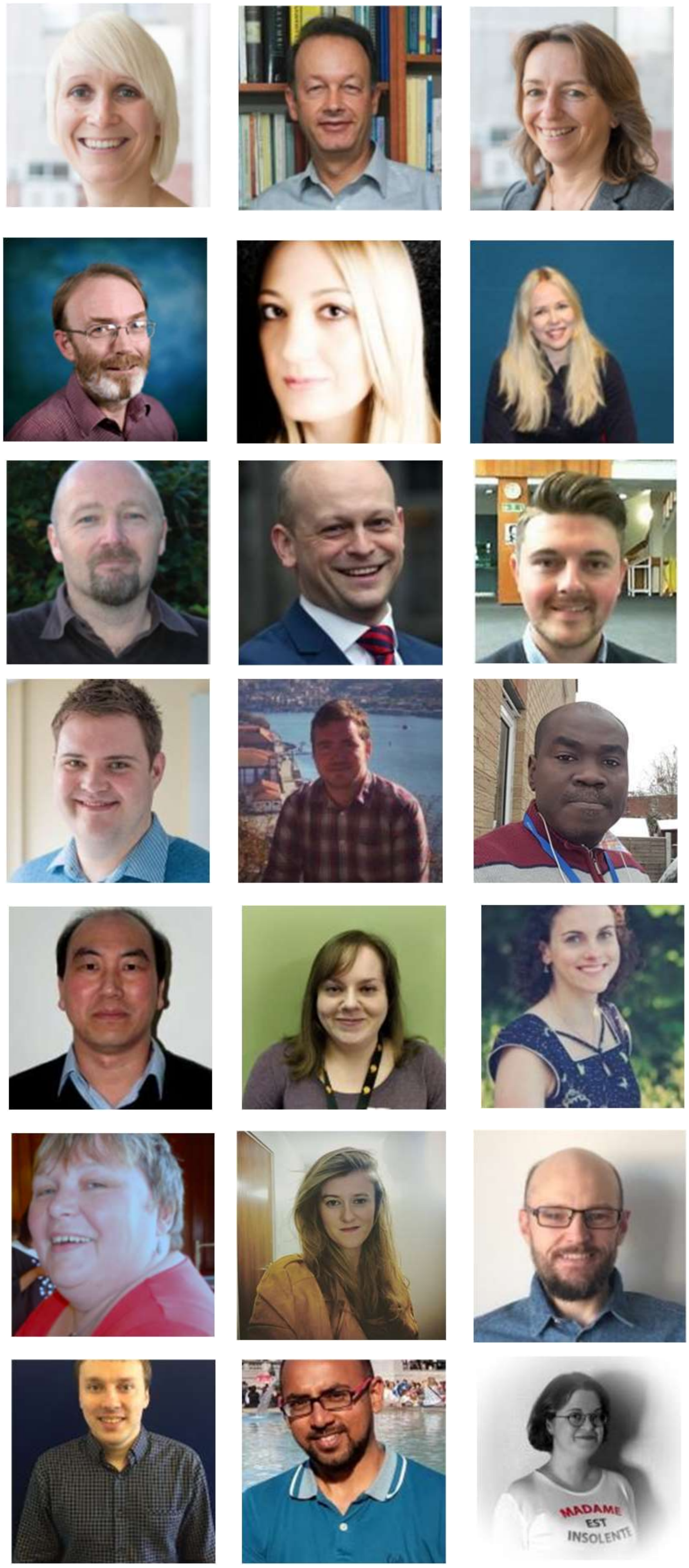
Find Out More

For more details about CorCenCC, associated project publications and satellite projects, visit the main website at www.corcencc.org

Video user guides are available at www.youtube.com/@corcencc

You can also follow us on Twitter @CorCenCC

To contact the team, email CorCenCC@cardiff.ac.uk



CorCenCC was an ESRC/AHRC funded research project (Grant Number ES/M011348/1). The team included **PI** - Dawn Knight; **CIs -** Tess Fitzpatrick, Steve Morris, Irena Spasić, Paul Rayson, Enlli Thomas, Mark Stonelake, Alex Lovell, Jonathan Morris and Jeremy Evas; **RAs** - Steven Neale, Jennifer Needs, Mair Rees, Scott Piao, Ignatius Ezeani and Gareth Watkins; **PhD students** - Vigneshwaran Muralidaran and Bethan Tovey-Walsh; **Consultants** - Kevin Donnelly, Kevin Scannell, Laurence Anthony, Tom Cobb, Michael McCarthy and Margaret Deuchar; **Project Advisory Group** - Colin Williams, Karen Corrigan, Llion Jones, Maggie Tallerman, Mair Parry-Jones, Gwen Awbery, Emyr Davies, Gareth Morlais, Owain Roberts, Aran Jones and Andrew Hawke.