

1. If there's anything to add to the chair's introduction of me, do so now

Developing the Cornish Dictionary using open source tools and data

Davydh Trethewey

Project Support Assistant, Sodhva Kernewek - Cornish Language Office
Konsel Kernow - Cornwall Council
davidtreth@gmail.com
taklowkernewek.neocities.org

Keskusulyans Wikimedia Kolm Keltek - Wikimedia Celtic Knot
conference,
5th July 2019, Penryn Campus

Previous Online Dictionary

- Standard Written Form of Cornish online dictionary
cornishdictionary.org.uk (Internet Archive Wayback Machine)

cornish language partnership
maga cornish dictionary / gerlyver kernewek

Welcome / Dynnargh

Cornish words

[Apply](#)

English words (exact)

[Apply](#)

English words (include)

[Apply](#)

Search

[Apply](#)

Cornish / Kernewek A-Z
English / Sonesek A-Z
Kampolla / Feedback

Abbreviations /
Bertheansow

'Middle' and 'Late' Cornish
forms / Fomys 'Koe' ha
'Divedhes'

Pronunciation / Leveryans

Traditional graphs / Grafys
hengovek

The team / Kesoberoryon

login

Welcome to the new online, searchable, dictionary of Cornish. The dictionary is written in the Standard Written Form of Cornish and the intention is that it will be constantly updated and extended. As well as adding new terminology and extending the range of the dictionary, over time information about each word and usage will also be made available as will audio files to allow the pronunciations to be heard.

A pdf version is available on the MAGA website and this will be updated on a 6 monthly basis to reflect additions to the online version.

Feedback on this resource is most welcome and a feedback form is included for your use.

MAGA hopes that this will be a useful addition to the wide range of resources now available to help both speakers and learners of Cornish. For further information about Cornish, resources and events, visit the website www.magakernow.org.uk

or contact the office on cornishlanguage@cornwall.gov.uk

Dynnargh dhe'n gerlyver hwiadow warlinen nowyth a Gemewek. Skrifys yw an gerlyver y'n Fuvr Savonek Skrifys a Gemewek ha'n mynnas yw y fydh prest nowythys hag ydyrnys. Kefrys ha newons termynologeth nowyth hag ydyrna etander an gerlyver, dres termyn kedhow a-dro dhe bub ger ha'y usadow a vydh kavadow, yn herydh restrerow sonek dhe asa bos klewys an leveryans.

Vershyon pdf yw kavadow war wlaova MAGA hag y fydh herma nowythys pub 6 mis dhe dhasowynys keworansow dhe'n versyon warlinen.

Dasle war an asroth ma a vydh meur dhynerhs ha furven dhasle yw komprehendys rag agas us.

MAGA a wylt y fydh herma keworans dhe les dhe'n asrothow a les kinda lemmyn kavadow dhe ri gweres ha keworion ha dyskoryon a Gemewek. Rag kedhow pella a-dro dhe Gemewek, asrothow ha keworow, vrysya an wlaova www.magakernow.org.uk

po kevada orth an sodhva war cornishlanguage@cornwall.gov.uk

NEWS
[SWF Review](#)

This dictionary has now been updated following the SWF Review. The PDF version will be available shortly. We would be grateful if users would report any accidental errors using the feedback form....
[read more](#)

1. This was the previous website up to June 2019
2. Different search boxes for English-Cornish and Cornish-English

Features in an ideal online dictionary

- Improve usability on different platforms desktop/tablet/mobile
- Cater for the various users of the language, including users of different varieties of Cornish
- Show personal forms for verbs and prepositions
- Ability to add sound samples
- Disambiguation of translation equivalents

1. Previous version could be cumbersome to use on mobile platforms. Maes T uses Google Web Toolkit allowing a responsive web-based interface. Maes T allows the backend data to be separated from the frontend website, for Welsh this has allowed use on different websites and mobile apps using an API. The Language Technologies Unit in Bangor has created a Wordpress plugin *Porthydd* to assist its deployment.^[4]
2. Note difference between Middle and Late forms.
needs of beginning learners, advanced learners, fluent speakers, Akademi members doing language development work.
3. Personal forms were done for prepositions, which combine with personal pronouns in Cornish to form inflected forms.
hasn't been done for verbs, since this is a larger set of words, several tenses for each verb. perhaps show the most common paradigms and the auxiliary verbs?
4. Having sound samples is probably better than using IPA
beginners may not understand these symbols. some of them aren't really agreed on, it is a Standard Written Form and may not be prescriptive for pronunciation.
5. e.g. [understanding](#) can mean an agreement, or knowledge. note dictionary also matches verb form understand, this is a feature of Maes T.

Things to improve

- English to Cornish, and Cornish to English had been separately created in previous version
- Some errors and inconsistencies
- Updates according to 2013 review of the Standard Written Form
- Integration of work done by Terminology Panel of Akademi Kernewek
- Provide platform for further work by Akademi Kernewek

1. Previous version had been done in a software called TshwaneLex.
2. Some review updates had already been done in the previous software, but not all.
3. Terms standardised by the Terminology Panel and recommended for use.
4. Akademi has a process for proposing a term, and wanted to be able to place it online but marked as a 'proposed term' so that the community would be able to comment on it before it became an established word in the language. There is a similar process in Maes T for Welsh, but this is within the group of domain experts prior to online publication.

Dictionary data

- Exported from the software used in the previous version as an XML file
- Each word or phrase in the dictionary is a <lemma> tag group
- Various information in sub-tags
- e.g. pronunciation, part of speech, plural, English glosses, example sentence, etymology, attestations in the traditional texts

1. XML is eXtensible Markup Language, which works along a similar principle to Hyper Text Markup Language, which web pages are written in, but being extensible, any text can be used as tag names.

Tidying the XML

- Use Python [Beautiful Soup](#) to analyse the XML
- Allows any errors or inconsistencies to be spotted by looping through the <lemma>s in the dictionary
- Simplify some of the structure, move subentries to their own <lemma>s
- Collaboration with Dewi Bryn Jones (Bangor University) to enable it to be in a format suitable for import into Maes T

1. Certain words / phrases were in <partofspeechgroup> and <subentry> tags within <lemma>s. this could lead to duplication where one phrase could exist as subentries of more than one <lemma>
2. Check cases where no part of speech is defined etc.
3. Count occurrences of descendant tags of <lemma>.
4. Count how many <lemma> have each part of speech.

Variants within the language

- Cornish as a revived language derives from sources at different time epochs
- Broadly speaking, two time periods of Middle and Late Cornish
- Different groups within the revival have based the revived language primarily on Middle or Late sources
- Orthographic decisions have sometimes made these seem further apart than they really are

1. One of the major data processes undertaken by DT was to combine each word into a single <lemma> and define the M and L variants where they exist within this.
2. Previously they lived in independent <lemma> and if you searched the M version it relied on references tags having been added to refer to the L version.
3. The previous data was not always systematic about how M / L distinction was marked. The M and L forms were independent of each other.

Example of XML with M/L distinction

aswa

```
<lemma>
<lemma.middlelemmasign>
<middlespelling>aswa</middlespelling>
</lemma.middlelemmasign>
<lemma.latelemmasign>
<latespelling>ajwa</latespelling>
</lemma.latelemmasign>
<lemma.middlepronunciation>['azwa]</lemma.middlepronunciation>
<lemma.latepronunciation>['æɟ(w)ɐ]</lemma.latepronunciation>
<lemma.partofspeech>n.f</lemma.partofspeech>
<lemma.middleplural>aswaow</lemma.middleplural>
<lemma.lateplural>ajwaow</lemma.lateplural>
<sense><te><te.te>breach</te.te></te></sense>
<sense><te><te.te>gap</te.te></te></sense>
</lemma>
```

1. On the website you can click through to the L version and vice versa, and to the English glosses.

Example of XML with no M/L distinction

a-ugh

```
<lemma>
<lemma.lemmasign>
<spelling>a-ugh</spelling>
<homonymnumber>(1)</homonymnumber>
</lemma.lemmasign>
<lemma.middlepronunciation>[a'y:x]</lemma.middlepronunciation>
<lemma.latepronunciation>[ə'ɪʊʰ]</lemma.latepronunciation>
<lemma.partofspeech>adv</lemma.partofspeech>
<sense><te><te.te>above</te.te></te></sense>
</lemma>
```

1. In this case there is only one spelling which is used by both Middle and Late Cornish.
There is however a <homonymnumber> tag because a-ugh can also be a preposition, which is listed as a separate <lemma>.

Example of XML - a-ugh (2)

a-ugh

```
<lemma>
<lemma.lemmasign>
<spelling>a-ugh</spelling>
<homonymnumber>(2)</homonymnumber>
</lemma.lemmasign>
<lemma.partofspeech>prp</lemma.partofspeech>
<sense><te><te.te>above</te.te></te></sense>
<sense><te><te.te>over</te.te></te></sense>
<lemma.personal.forms>
<sg1p>a-ughov</sg1p>
<sg2p>a-ughos</sg2p>
<sg3pm>a-ughto</sg3pm>
<sg3pf>a-ughti</sg3pf>
<pl1p>a-ughon</pl1p>
<pl2p>a-ughowgh</pl2p>
<pl3p>a-ughta</pl3p>
</lemma.personal.forms>
<lemma.late.personal.forms>...</lemma.late.personal.forms>
</lemma>
```

1. The preposition homonym has personal forms listed.
2. There is also a <lemma.late.personal.forms> tag group which was edited out on the slide for brevity containing the Late personal forms.

Maes T software

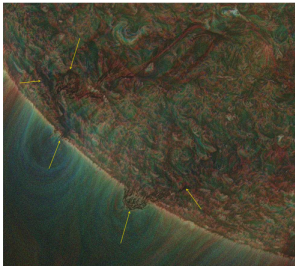
- Software that is used for terminology dictionaries in Welsh online at <http://termau.cymru>
- Developed by [Language Technologies Unit](#), Canolfan Bedwyr, Bangor University [1] [2]
- Deployable via an API to the web or apps [4]
- Maes T has also been used for geiriadur.bangor.ac.uk and www.termiaduraddysg.org where sound samples are also presented

1. Maes T was guided by ISO standard formats and practices for terminology and lexicography. The ISO standard 15188 recommends a working group of 5-8 subject specialists and a terminologist. Provide interface designed to be user-friendly to non-terminologists, bilingual English-Welsh.
2. There is a four stage process
 1. Collecting Terms - record candidate terms for the source language term.
 2. Defining the concept - define in the source and target languages, add a disambiguator where needed.
 3. Standardizing Terms -record normative status of candidate terms.
 4. Linguistic Information - record part of speech, gender, plural forms etc.
3. Data can be served via an API, features such as showing related entries, popups to explain pos tags. matching images from Wikidata, Wikimedia Commons, linking to Wikipedia.
4. Potential benefits of ISO standard based approach of Maes T for Cornish, such as software locali[sz]ations.

Termau.cymru

filament *(astronomy)* ffilament **eg** ffilamentau *(seryddiaeth)*

Tafod o nwy dwys cymharol oer wedi ei ìoneiddio (~10,000K), yn gaeth mewn bwndeli cymhleth o faes magnetig yn atmosffer isel yr Haul. Mae ffilamentau'n ymddangos yn dywyll yn erbyn cryfder yr Haul y tu ôl iddynt.



A tongue of dense relatively cool ionized gas (~10,000K), held in place by complex bundles of magnetic field in the Sun's low atmosphere. Filaments appear dark against the brightness of the Sun behind them.

Gelriadur Termau'r Coleg Cymraeg Cenedlaethol - Mathemateg a Ffiseg

termau.cymru/#filament [3]

1. Filament: termau.cymru/#filament. In this case, the image was contributed by an expert, along with a definition.

Termau.cymru

five-spot ladybird *Coccinella quinquepunctata* buwch gota bum smotyn eb buchod cwta pum smotyn



Buchod Cwta. Cymdeithas Edward Llwyd 2014

termau.cymru/#ladybird

1. Ladybirds: termau.cymru/#ladybird. In this case, Wikimedia commons images were used (manually chosen during a project with *Cymdeithas Edward Llwyd*).

Termau.cymru

herring gull *Larus argentatus* gwylan penwaig **eb** gwylanod penwaig

[gwylan penwaig ar Wikipedia](#)



Adar y Byd. Cymdeithas Edward Llwyd a Chymdeithas Ted Breeze-Jones 2015

[termau.cymru/#gull](#)

1. This was added as part of *Adar Y Byd*, a dictionary of bird names from *Cymdeithas Edward Llwyd*. This had 9500 entries, so an automated method was used.
2. There is a species of moth called 'Silurian' and this found a Wikimedia image of a Doctor Who character. [3]
Note also: [termau.cymru/#apple](#) which showed as of July 2019 a picture of Apple headquarters rather than the fruit.
[xkcd.com/2140 Reinvent the Wheel](#)
Also beware of possible API changes.

Adapting Maes T to better serve Cornish

- Collaboration with Dewi Bryn Jones to adapt the Maes T software for Cornish
- Some relevant grammatical differences between Welsh and Cornish
- Other changes come from using it for a general dictionary website rather than terminology dictionaries that usually have a 1:1 correspondence between Welsh-English in a given context

1. The original purpose of Maes T was to support terminology dictionaries which aimed at a one to one mapping between source language and target language terms. Although it had been used since for more general dictionaries such as [Geiriadur Bangor](#).
2. Welsh grammars tend to treat the singulative / collective nouns as a normal singular / plural distinction e.g. coeden - coed, plentyn - plant, where the noun loses an ending to become plural.
3. The plural of the singulative is a feature of (at least some forms) of Cornish, for a number of individual items, rather than en masse, for at least some n.coll. E.g. gwydh, gwedhen, gwydhennow. Not currently shown in dictionary.

Transition to editing within Maes T

- Once the structure is stable, move away from manually editing XML to editing within Maes T
- More practical for a wider range of people e.g. Dictionary Panel members of the Akademi Kernewek to edit
- Manually editing an XML file can be error prone, which was mitigated by the Python scripts validating / analysing it

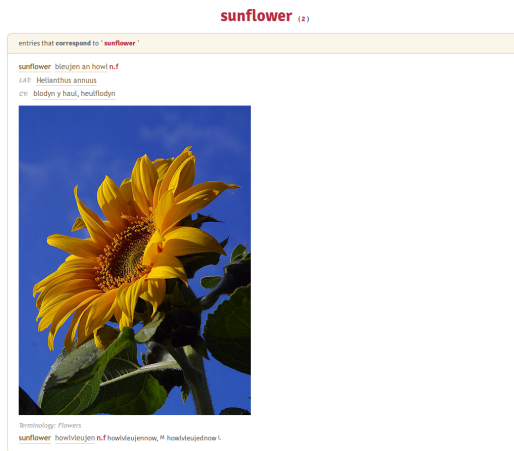
1. In production, the data would be edited by a number of different people who may be non-programmers.
2. Maes T can provide a platform for this in a similar way as to the Welsh terminology development.

Terminology Panel

- [Akademi Kernewek](#), the Cornish language academy has a [terminology panel](#) to research new terms for the language
- A number of subject areas have been considered so far: plants, insects, mining, minerals, architecture, grammar

1. Briefly outline process by which Akademi intends to do this, and terminology areas in development.
2. [Akademi terminology policy](#) outlining process of creating a term.
3. [Akademi dictionary policy](#) according to which a terminology item is recommended and after a period of review becomes considered part of the main dictionary

Using open source data from Wikimedia



www.cornishdictionary.org.uk/#sunflower

1. Explain a bit about linking to Wikimedia Commons images of plants, and to Wikipedia.
2. See [3] for more information on [Porth Termaw Cenedlaethol Cymru](#) linking to Wikimedia images. This included automatically finding them using English and Latin names.

New dictionary website

- Demonstrate the new dictionary website (demo of cornishdictionary.org.uk)

1. New version is now at the main URL.
2. Possibly do the demo at the end or in an unconference session subject to time, maybe better for flow of the talk to move to conclusions now.
3. Responsive design - e.g. cornishdictionary.org.uk/#moon automatically switches to two-column layout where there are matches in both languages, and if the screen width is low, these are stacked vertically rather than side-by-side.

Conclusions and future ideas

- We already have some example sentences, but could have many more of these, and audio of them by speakers
- Method of handling Middle and Late variants allows multiple variants to be supported while keeping them semantically as one <lemma> item

1. Other features that would be nice?
2. Make plurals / singulatives searchable e.g. mergh, gwedhen search aware of mutation such that 'gath' finds 'kath', awareness of 'traditional' spellings that some users prefer such that searching 'cath' finds 'kath' spellings from other forms of Cornish e.g. gwydhenn (Kemmyrn) - gwedhen (SWF) could M / L, and/or the 'traditional' forms be offered as a user selection at the front-end? and a choice between a strict and 'fuzzy' search?
3. Maes T does this for Welsh, using lemmatizers, spelling and grammar checkers. e.g. geiriadur.bangor.ac.uk/#plant lemmatizes 'plant' to the singular 'plentyrn' "child". It will also present the English entries for 'plant'. geiriadur.bangor.ac.uk/#mor gives the conjunction/adverb mor and the noun môr, as well as bôr (an alternate spelling for bore, b could have nasal mutated to m). Maes T customizes results using different API keys including different terminology resources and search settings.
4. Possibilities of wider linking to Wikimedia items including Wikidata?
5. Linkage with place-name map?



Tegau Andrews and Gruffudd Prys. "Terminology Standardization in Education and the Construction of Resources: The Welsh Experience". In: *Education Sciences* 6.1 (2016), p. 2.



Tegau Andrews, Gruffudd Prys, and Dewi Bryn Jones. "The Maes T System and its use in the Welsh-Medium Higher Education Terminology Project". In: *Creation, Harmonization and Application of Terminology Resources* (2011), p. 49.



Gruffudd Prys et al. "Crossing between environments: the relationship between terminological dictionaries and Wikipedia". In: *Terminologie(s) et traduction*. Peter Lang, 2018. ISBN: 9783631746431.























Gruffudd Prys et al. "Distributing Terminology Resources Online: Multiple Outlet and Centralized Outlet Distribution Models in Wales". In: *The 2nd Workshop on the Creation; Harmonization and Application of Terminology Resources*. 72. Linköping University Electronic Press. 2012, pp. 37–40.

Possible things to talk about in unconference sessions

- Another way of doing things would be to generate static HTML pages programatically from the XML, which I also did
- As a side project of this, I programatically matched the English glosses to Unicode character names

Cornish Emojis

| | | |
|---|------------------------|--------------------|
| kasek | | mare |
|  | | |
| kasorek | | militant, military |
|   | | |
| kastel | | castle, hill fort |
|    | | |
| kath | | cat |
|  찰            | | |
| kav | PHAISTOS DISC SIGN CAT | cave |
|  | | |
| kavas | | tin |
|  4. | | |

Cat Emoji (circa 1500BC) from the Phaistos Disc