

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 / Dynmargh

Cornish words

[Apply](#)

English words (exact)

[Apply](#)

English words (include)

[Apply](#)

Search

[Apply](#)

Cornish / Kernewek A-Z

English / Sowsnek A-Z

Karppola / Feedback

Abbreviations /
Berhewansow

'Middle' and 'Late' Cornish
forms / Fomya 'Kies' ha
Owedher

Pronunciation / Leveryans

Traditional graphs / Grafys
hengovek

The team / Kesoberoryon

login

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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>
```

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>
```

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>
```

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

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

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

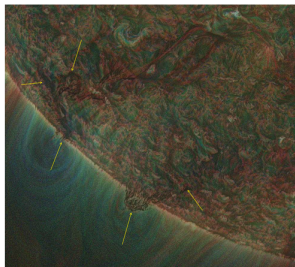
Maes T software

- 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

Termau.cymru

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

Tafod o nwy dwys cymharol oer wedi ei ÷oneiddio (~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.

Geiriadur Termau'r Coleg Cymraeg Cenedlaethol - Mathemateg a Ffiseg

[termau.cymru/#filament](#) [3]

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

Termau.cymru

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

[gwylan penwaig ar Wicipedia](#)



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

[termau.cymru/#gull](#)

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

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

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

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

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

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

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

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

Using open source data from Wikimedia

sunflower (2)

entries that correspond to 'sunflower'

sunflower bleujen an howl *n.f*
LAR: Helianthus annuus
CK: blodyn y haul, heufloedyn



Terminology: Flowers
sunflower howlbleujen *n.f* howlbleujennow, H howlbleujednow L

www.cornishdictionary.org.uk/#sunflower

New dictionary website

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

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

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

References



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
kav 	<div>PHAISTOS DISC SIGN CAT</div> cave
kavas  	tin

Cat Emoji (circa 1500BC) from the Phaistos Disc

Cornish Emoji Dictionary taklowkernewek.neocities.org/emojiskernewek