

Creating Software Tools for Cornish with Python

David Trethewey

davidtreth@gmail.com
taklowkernewek.neocities.org

Cornish Language Research Network (Skians),
30th September 2016, Tremough Campus

Language Technology for Cornish

- SWF online dictionary cornishdictionary.org.uk
- Glosbe - the multilingual online dictionary glosbe.com/kw
- Gerlyver Kernewek-Kembrek (by Dr Paul Bowden + Dr Kevin Donnelly).
Online Cornish-Welsh dictionary with 4000 words.
Machine translation Cornish → English program *kern* by Paul Bowden. kevindonnelly.org.uk/kernewek
- Transliteration software to SWF by Steve Harris and Peter Harvey.

Python Natural Language Processing Toolkit (NLTK)



Image stolen from update.hanser-fachbuch.de/2013/09/artikelreihe-python-3-nltk-natural-language-toolkit

Descriptive Corpus Statistics

- Traditional Cornish texts in computer readable form howlsedhes.co.uk
- Some modern texts from www.kernewegva.com and www.learncornishlanguage.co.uk.

Corpus analysis - word frequencies

The screenshot shows the 'TaklowKernewek Tools' interface. On the left, a sidebar titled 'Tekst' lists various text files, with 'Bewnans Meryasek' selected. The main area is titled 'Dewis Gwrythyans' and contains several analysis options and filters. The 'Text: Bewnans Meryasek' section displays the following statistics:

- number of words = 26819
- number of different words = 4667
- Lengths of words in descending order of frequency: [(1, 3080), (2, 4813), (3, 5094), (4, 3857), (5, 3271), (6, 2637), (7, 1697), (8, 1180), (9, 612), (10, 385), (11, 115), (12, 57), (13, 18), (14, 2), (18, 1)]
- Top 15 words: ['a', 'y', 'n', 'dhe', 'ha', 'yn', 'an', 'ow', 'my', 'yw', 'c', 'ny', 'na', 're', 's']
- Top 15 words of 5 or more letters: ['meryasek', 'krist', 'arloedh', 'lemmyn', 'maria', 'meriadcus', 'dhymm', 'sertan', 'dhymmo', 'finit', 'episcopus', 'primus', 'comes', 'secundus', 'yredi']

Below the statistics, there are input fields for 'Keworowgh isella niver a lytherennow rag raiyow menowghder ger a-woeles:' (set to 5), 'Keworowgh niver a eryow dhe dherivas an menowghder a-woeles: defowt = 20' (set to 15), and 'Keworowgh ger dhe geheveli menowghderow dres an tekstow:'. There are also buttons for 'Keworra ger dhe'n rol' and 'Klerhe an rol'. At the bottom, there are buttons for 'Kopi dhe'n Klyppbordh', 'Dalleth', 'Klerhe Tresennow', and 'Kwitya'.

The most common words in Bewnans Meriasek, and the most common of 5 of more letters.

Live demo!

- Demonstration of corpus analysis module from TaklowKernewek tools.

Mutation

The screenshot shows a window titled "Mutatya" with a sidebar on the left and a main content area on the right. The sidebar is titled "Studh Treylyans" and contains a list of radio buttons for selecting a study type: 1 (heb treylyans), 2 (medhel), 3 (hwythys), 4 (kales), 5 (kemmyskys), 6 (kemmyskys wosa 'th), 7 (kildreylyans), Lytherennans, and hengovek (SWF/T). The main content area is titled "Gorrewgh ger kernewek a-woles mar pleg" and features a text input field containing the word "garr". Below the input field, the text "Y halsa bos an ger:" is displayed in red. Further down, the text "Furv didreylys:" is followed by "garr" on the right, and "Treylyans medhel (studh 2) diworth: karr" is displayed below it. At the bottom of the window, there are three buttons: "Kopi dhe'n Klyppbordh", "Mutatya", and "Kwitya".

Using the input word *garr* the program shows that it could be an unmutated form, or a mutation of *karr*.

Numbers

The screenshot shows a window titled "Niverow" with the heading "Dewisyow". On the left, there are two checked options: "Usya Hanow" and "Hanow Benow". The main area displays three examples of inflected nouns:

- Example 1: "Gorrewgh niver a-woles mar pleg" with a text box containing "3".
- Example 2: "Gorrewgh hanow kernewek a-woles mar pleg" with a text box containing "kath".
- Example 3: "Gorrewgh hanow liesplek a-woles mar nag yw -ow" with a text box containing "kathes".

Below these examples, the word "teyr hath" is highlighted in red. At the bottom, there are four buttons: "Klerhe", "Kopi dhe'n Klyppbordh", "Diskwedh Niver", and "Kwitya".

A number and a noun in Cornish. It is necessary to tell the program whether to use the noun, and if it is feminine.

Numbers

The screenshot shows a window titled "Niverow" with a header "Dewisyow". On the left, there are two checked checkboxes: "Usya Hanow" and "Hanow Benow". The main area displays three examples of number inflection:

- Example 1: "Gorrewgh niver a-woles mar pleg" with a text box containing "343".
- Example 2: "Gorrewgh hanow kernewek a-woles mar pleg" with a text box containing "kath".
- Example 3: "Gorrewgh hanow liesplek a-woles mar nag yw -ow" with a text box containing "kathes".

Below these examples, a yellow highlighted box contains the text: **tri hans, tri ha dew ugens a gathes**. At the bottom, there are four buttons: "Klerhe", "Kopi dhe'n Klyppbordh", "Diskwedh Niver", and "Kwitya".

A number and a noun in Cornish. For a number with more than three elements, it follows the number + a^2 + plural noun form.

Inflecting verbs

The screenshot shows a software window titled "Inflektya Verbow Kernewek". The main area is titled "Dewisyow" (Options) and "Amser" (Time). It features several panels for selecting options:

- Person:** A list of pronouns including My, Ty, Ev, Hi, Ni, Hwi, I, Anpersonek, and Pub Person.
- Raghenwyn a syw:** Options for "Heb raghenwyn a syw", "Raghenwyn a syw", and "Raghenwyn a syw gans poeslev".
- Usya FSS?:** A checkbox for "Ynworrans + eskorrans FSS".

The main display area is titled "Gorrewgh verb kernewek a-woeles mar pleg:" and contains a text input field with "gweles". Below the input is a table showing the conjugation of the verb "gweles" for various persons:

Anpersonek	:	gwelir
My	:	gwelav vy
Ty	:	gwelydh jy
Ev	:	gwel ev
Hi	:	gwel hi
Ni	:	gwelyn ni
Hwi	:	gwelowgh hwi
I	:	gwelons i

At the bottom of the window, there are buttons for "Klerhe", "Kopi dhe'n Klyppborth", "Inflektya Verb", and "Kwitya".

Inflecting the regular verb *gweles* (to see).

Syllable segmentation

- Works via regular expressions in Python.
- Scans through input words and identifies number of syllables.
- Finds structure of syllable and which should be stressed.

Syllable segmentation

The screenshot shows a window titled 'Syllabenn Ranna Kernewek' with a sidebar on the left containing radio buttons for 'Mode Hir', 'Mode Berr', 'Mode Linenn', 'Rannans war-rag', and 'Rannans war-dhelegh'. The main area is titled 'Gorrewgh tekst kernewek a-woeles mar pleg:' and contains the text 'Dohajydh da.' Below this, a yellow-highlighted box displays the following analysis:

```
An ger yw: Dohajydh
Niver a syllabennow yw: 3
Hag yns i: ['Do', 'ha', 'jydh']
S1: Do, CV, hirder = [1, 1], hirder kowal = 2
S2: ha, CV, hirder = [1, 1], hirder kowal = 2
S3: JYDH, CVC, hirder = [1, 2, 1], hirder kowal = 4
Hirder ger kowal = 8

An ger yw: da
Niver a syllabennow yw: 1
Hag yns i: ['da']
S1: DA, CV, hirder = [1, 3], hirder kowal = 4
Hirder ger kowal = 4
```

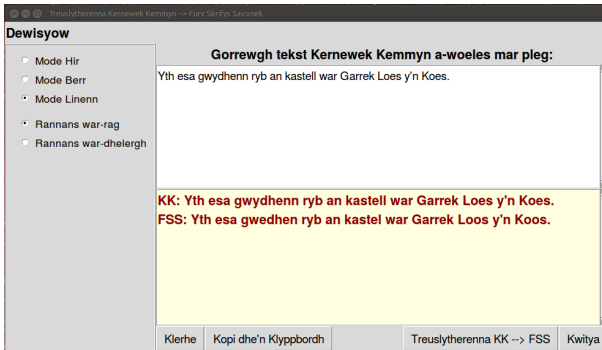
At the bottom of the window are buttons for 'Klerhe', 'Kopi dhe'n Klyppbordh', 'Diskwedh Syllabennow', and 'Kwitya'.

Long mode giving details of each syllable. The word *dohajydh* is among a list of words with unusual final stress.

Transliteration from KK to SWF

- Some substitutions such as *oe* → *oo* or *oe* → *o* depend on vowel length or syllable stress.
- Two steps, syllable level and word level substitutions.
- List of exceptions to general rules in a data file.

Transliteration KK → SWF



Line mode shows each line of the input interlinearly, Kernewek Kemmyn and SWF.

What is translation memory?

- Match same sentences or segments in a bilingual corpus.
- Assists translators by using previous experience in translating similar texts.
- Various proprietary and open-source software is available.
- Wikipedia: [Comparison of computer-assisted translation tools](#)
- Can save labour, and improve consistency.

A simple translation memory with Python NLTK

- Use NLTKs bigram and trigram finding functions.
- Bilingual corpus based on *Skeul an Yeth 1* example sentences.
- Option to ignore trivial bigrams like “in the” which are all stopwords (a list of common words defined in a NLTK corpus).

- Example input sentence is “Snowdon is the highest mountain in Belarus and Wales.”
- There is 1 sentence with trigram matches - “Brown Willy is the highest mountain in Cornwall.”.
- In fact there is a 5-gram match, which the program returns as 3 trigram matches.
- There are other sentences with bigram matches for “the highest”.

The highest mountain

```
Listing N-grams with a minimum of 1 non-stopword each:  
Common trigrams:  
Bronn Wennili yw an ughella menydh      -- Brown Willy is the highest mountain  
yn Kernow.                               -- in Cornwall.  
  
(is the highest), (the highest mountain), (highest mountain in)  
  
Common bigrams:  
Ny yll ev esedha war an ughella         -- He cannot sit on the highest one.  
huni.                                     --  
  
(the highest)
```

The first bilingual sentence has 3 trigram matches, and the second a single bigram match.

Introduction to WordNet

- WordNet is a lexical database of English
wordnet.princeton.edu
- Nouns, verbs, adjectives and adverbs are grouped into *synsets* each expressing a distinct concept.
- Synsets are interlinked by conceptual-semantic and lexical relations.
- For example hypernyms and hyponyms are more general and more specific categories.
- E.g. *bed* is a hyponym of *furniture*, and *bunkbed* a hyponym of *bed*.

Finding synonyms with WordNet

- A program in TaklowKernewek allows input of an English sentence, for which each word is converted into a list of hyponyms of its hypernyms.
- These may be synonyms, or related concepts.
- word: hill
 - Synset('hill.n.01'): a local and well-defined elevation of the land
 - Synset('mound.n.04'): structure consisting of an artificial heap or bank usually of earth or stones
 - Synset('hill.n.03'): United States railroad tycoon (1838-1916)
 - Synset('hill.n.04'): risque English comedian (1925-1992)
 - Synset('mound.n.01'): (baseball) the slight elevation on which the pitcher stands
 - Synset('hill.v.01'): form into a hill

Finding synonyms with WordNet II

- Hypernyms

Synset('natural_elevation.n.01'): a raised or elevated geological formation

Synset('structure.n.01'): a thing constructed; a complex entity constructed of many parts

Synset('baseball_equipment.n.01'): equipment used in playing baseball

Synset('shape.v.02'): make something, usually for a specific function

Finding synonyms with WordNet III

- Inputting “What is the highest hill in Cornwall?” produces:
- “What is the highest mountain in Cornwall?” which matches to “Brown Willy is the highest mountain in Cornwall.” in the *Skeul an Yeth 1* corpus.
- Also “What is the highest stadium in Cornwall?” and even “What is the highest baseball in Cornwall?” among many others.

Conclusions and future ideas

- Code is available at Bitbucket repository at bitbucket.org/davidtreth/taklow-kernewek

Future work:

- Part of speech tagging?
- Translate to Javascript for web use?
- Games to assist learning?
- Ideas from the community of Cornish users please.

For Further Reading I



Python Natural Language Toolkit

www.nltk.org

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly publications.



Welsh National Language Technologies Portal

techiaith.cymru



Prof. Kevin Scannell's website containing a large number of links on language technologies for minority languages.

borel.slu.edu/nlp.html

For Further Reading II



Language Engineering Resources for the Indigenous Minority Languages of the British Isles and Ireland (Lancaster University)

includes a proposed part of speech tagset for Cornish by Jon Mills.




www.lancaster.ac.uk/fass/projects/biml



Publications by Dr. Jon Mills including papers about language technologies for Cornish.

[link to Dr. Jon Mills site on Academia.edu](#)

For Further Reading III

-  Giellatekno, the Center for Saami language technology, Arctic University of Norway.
giellatekno.uit.no/index.html
including some work on Cornish:
giellatekno.uit.no/cgi/index.cor.eng.html.
-  eSpeak - an open-source “formant synthesis” speech synthesis software package.
espeak.sourceforge.net
-  Apertium - a free/open-source machine translation platform.
www.apertium.org