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.
The most common words in Bewnans Meriasek, and the most common of 5 or more letters.
Live demo!

- Demonstration of corpus analysis module from TaklowKernewek tools.
Using the input word *garr* the program shows that it could be an unmutated form, or a mutation of *karr*. 
A number and a noun in Cornish. It is necessary to tell the program whether to use the noun, and if it is feminine.
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

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

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 NLTK's 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”.

Listing N-grams with a minimum of 1 non-stopword each:
Common trigrams:
Bonn Wenni lì 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*.
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

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](http://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](http://link-to-dr-jon-mills-site-on-academia.edu)
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