Abstract for Skians Conference
30th September-1st October 2016

David Trethewey
davidtreth@gmail.com
taklowkernewek.neocities.org

Gwrians toulys medhelweyth rag Kernewek gans Python

Y'n areth ma, my a wra displegya ow ober gul medhelweyth rag Kernewek yn Python. Usys yw an *Natural Language Processing Toolkit* www.nltk.org yn Python dhe amontya statystygyon korpus, ha skrifys o programmow Python dhe wul treylyansow kernewek, skrifa niverow kernewek, ranna tekst kernewek yn syllabennow, inflektya verbow kernewek, ha treuslytherenna tekst Kernewek Kemmyn bys dhe'n Furv Skrifys Savonek. Kodenn fenten an ober ma yw igor, ha kavadow diworth ow powesva Bitbucket (bitbucket.org/davidtreth/taklow-kernewek). Galladewderyow pella an displegyans a vydh styrys ha'gas tybyansow gelwys.

Creating Software Tools for Cornish with Python

In this presentation, I explain my work creating software for Cornish in Python. The *Natural Language Processing Toolkit* www.nltk.org in Python is used to calculate corpus statistics, and Python programs have been written to do mutation in Cornish, write out Cornish numbers, divide Cornish text into syllables, inflect Cornish verbs and transliterate Kernewek Kemmyn text to the Standard Written Form. The source code of this work is open, and available from my Bitbucket repository (bitbucket.org/davidtreth/taklow-kernewek). Potential future development will be discussed and your ideas invited.

Notes

Context and Background

Much of the interest in the Cornish language has centred around its importance for Cornwall's distinctive history, however it is also important if it is to continue to grow in usage in the 21st century as a revived language, to consider the need for development of language technologies and tools in respect of it, if only to keep up with the ways in which language is used today. I will be speaking primarily about my own work using Python, which contains a number of basic applications, and invite discussion about future directions of work.

Dictionaries

There are some examples of online dictionaries of Cornish, such as the SWF dictionary, and some independent work, including a Cornish-Welsh dictionary kevindonnelly.org.uk/kernewek, and the multilingual project Glosbe, which also includes some translation memory glosbe.com/kw/en. The Tatoeba sentence database tatoeba.org/eng/about includes a small number of Cornish sentences. The Panlex project panlex.org/index.shtml aims to be able to express any lexical concept in any language. Wiktionary kw.wiktionary.org.

Previous work

Steve Harris has previously produced a KK to SWF transliterator. This was for pre-2013 SWF and is not currently online. There has also been work by Peter Harvey for Unified Cornish to SWF transliteration. Dr. Jon Mills has suggested a part of speech tagset for Cornish www.lancaster.ac.uk/fass/projects/biml/bimls3repor and www.lancaster.ac.uk/fass/projects/biml/cornish_tags.htm.

Online access to text corpus

Many of the traditional texts, including the entire Middle Cornish corpus, are available for download as texts. Some works have scans of the manuscript itself. New Testament and Psalms translation has online web access www.bible.com/versions/1079 abk-an-testament-nowydh-han-salmow-2014.

Python Natural Language Toolkit

Why Python? It is a widely used programming language which has a large number of libraries available to do various tasks, including NLTK which has been specifically created for Natural Language processing tasks.

Descriptive Corpus Statistics

I mainly have focused on word frequency and distribution of word length so far. Could in future extend this to bigrams and trigrams. Could extend to consider numbers of syllables in words.

TaklowKernewek utilities

Mutation

Though the scope of the program `mutatya.py` doesn't decide when a word should mutate, it is fairly simple to implement the rules of Cornish mutation itself in a Python function. The function takes in the word, and the state to mutate it to as a number from 2-6, where 2=soft, 3=breathed, 4=hard, 5=mixed, 6=mixed after infixed pronoun 'th. The program also allows a reverse mutation, whereby it can be determined whether a word input could be a mutated form, though this doesn't check whether the hypothetical original form exists as a word in Cornish or whether the mutation is grammatically possible.

Numbers in Cornish

The program `niverow.py` aims to convert input integers and convert to a Cornish number in words. It is possible to include a noun, though it is necessary to manually tell the program if it is feminine, and the plural form of the noun (used for large or complex numerals of more than 3 elements). Floating point numbers are processed one digit at a time.

Calendar app

The program `termynGUI.py` is a fairly simple extension of the numerals program, making use of the Python `time` library to write the date and/or time in Cornish, or give a greeting appropriate to the time of day.

Text to speech

Cornish text is preformatted to partially conform to Welsh orthographic conventions with a series of string replaces, that allows text to be processed by the software `espeak`. This program currently only works on Linux based systems that have espeak available on the command-line. It would be also possible to create a Cornish voice that could process Cornish text directly. A good speech engine would be more complex however, and speech recognition more so.

Inflecting verbs

The program `inflektya.py` inflects Cornish verbs for person and tense. Quite a lot of manual data entry was required for this, to create rules to treat different classes of words in different ways, and to specify the irregular verbs. The Grammar of Modern Cornish (3rd ed. 2000, Wella Brown) and Cornish Verbs (3rd ed. 2010, Ray Edwards/Kesva an Taves Kernewek) were used as the primary sources.

Syllable analysis

Syllable segmentation

Using regular expressions in Python, scan through input words, and match the regular expressions to syllables. This can be used to identify the number of syllables in a word, and identify whether a syllable is stressed, and its internal structure (CVC, VC, CV, V). Again a substantial amount of data entry to specify exceptions to general rules. Long mode details internal structure of each syllable. Short mode shows only the number of syllables of each word, and line mode is similar to short mode except the total number of syllables in a line is also output. It is possible to choose whether to segment forwards or backwards. Syllable boundaries can be a contested matter in linguistics, even among experts, so this program will not always be correct.

Transliteration KK to SWF

This builds on the syllable segmentation program, since some changes from KK to SWF depend on factors of vowel length or syllable stress, which are determined by the syllable's position in the word. There are two stages, firstly the syllable level substitutions are done and then word level changes. Exceptions to general rules are coded in a file. Long mode includes all of the syllable length and details information, short mode only the SWF text, and line mode KK and SWF interlinearly. This is not currently reversible, and may be difficult to do so, since generally KK is more specific than SWF, that is it may not be easy without context for a computer program to know that a SWF -o- comes from KK -oe- or not, for example.

Translation Memory

I don't have personal experience with the software packages typically used for translation memory, so do not discuss them here. My own work is based on bigram and trigram matching to the example sentences in Skeul an Yeth 1. The scope of this is initially English to Cornish translation.

Bigram and Trigram matching

Using NLTK tooks for bigram and trigram finding. The program presents trigram matches first, followed by bigrams. Within that, sentences with more matches are ranked higher and appear first. It is possible to choose whether to exclude N-grams consisting only of NLTK stopwords (a list of common English words).

Experimental work with WordNet

WordNet allows words with semantic similarity to be considered for bigram and trigram matching. This may assist with having a very small bilingual corpus available. The current program however generates a lot of false positives, such as the word 'hill' having one of its meanings as being a baseball term, and going one level up the synset hierarchy and going back down again includes a large number of baseball terms.

Summary

One of the biggest challenges is not having an enormous corpus available. Building the corpus is more important than specific choices of software. Multilingual projects like PanLex, and Glosbe are interesting avenues.

Future directions

Part of speech tagging could be a useful extension, which would also be a foundation for other future applications, such as analyis of grammar, a spellchecker, and ultimately machine translation. It would likely be necessary to produce a manually tagged corpus, and before that to decide on a tagset to use. Web development - much of the above could be done in Javascript for web applications.

Notes on further reading

The eSpeak voice for Welsh, is described as a fairly basic implementation so it should be possible to improve upon it by producing a voice file that directly handles Cornish.

The Apertium open-source software provides a platform for machine translation, using a rule-based morphological analysis approach.