On Simplifying the Lexical Tagging of Cornish Texts

Ken George

Bosprenn, Keveral Lane,
Seaton, Torpoint,
Cornwall, G.B.

kjgeorge@plymouth.ac.uk Email
+44 1 752 232461 Tel
+44 1 752 232406 Fax

Abstract

Work has begun on a new computer-aided analysis of the Cornish texts, using ***Kernewek Kemmyn*** as a standard comparison instead of Unified Cornish. Suffixed and mutated words need to be related to a head-word, but instead of tagging every such word, they are identified using the principles of relational data-bases; only homographs in the standard text need be tagged. Details of this labour-saving procedure are given.

Introduction

In the discussion about the CELT project at XI-ICCS in Cork in 1999, the remark was made that metatextual grammatical tags had been added to scarcely any Celtic texts, because of the volume of work involved. The work may be prohibitively onerous for syntactic tagging, but for lexical tagging it is not. The labour may be much reduced, because there is no need to tag every mutated or suffixed word in a text. Instead, only the homographs need to be tagged, to differentiate them, and recourse may be had to the principles of relational databases to deal with the rest of the exercise.

Cornish was spoken traditionally as a vernacular until the end of the eighteenth century, and was revived in the twentieth century. The first computer-aided analysis of traditional Cornish texts was carried out by the present author (George, 1988), resulting in a large data-base, used primarily in order to study the phonological history of Cornish (George, 1984). Because the spelling of traditional Cornish was not fixed, it was helpful to have a standardized orthography to which the textual spellings could be related. Nance's Unified Cornish (1929) was used for this purpose; the analysis showed, however, that Unified Cornish could be improved (George, 1986), and a new orthography, known as ***Kernewek Kemmyn***, was introduced to replace it.

The data-base was later used in the compilation of a new dictionary using ***Kernewek Kemmyn*** (George, 1993). In this dictionary, a tripartite authentication code was devised and given for each head-word. This included an indication of where and how often a given head-word was attested. It was the first time that such detailed information had been provided in a Cornish dictionary, and the idea received widespread acclaim. Nevertheless, the authentication codes were not always complete or correct, as pointed out by Edwards (1999). The relevant files have been edited in order to improve the authentication codes, in anticipation of a new edition of the dictionary. Rather than continue with this piecemeal approach, however, the time has come to carry out another detailed analysis of the texts.

Nessa Tremen – a new analysis of the Cornish texts

The new computer-aided analysis is known as ***Nessa Tremen*** 'Second Pass', the first pass being the analysis carried out in the early 1980s. George (1988) noted that "If the work were to be done again, it would be better to use a phonemic orthography as a comparison standard". ***Kernewek Kemmyn*** is not completely phonemic, but is much more so than Unified Cornish. Dunbar and George (1997, p. 11) described the idea of ***Nessa Tremen*** as "using an iterative technique to produce an ever more accurate solution".

Since the first analysis, computers have become much more powerful, large and common. Whereas it was previously necessary to process texts line by line, reading and writing data sequentially to work-files, it is now possible to read the entire extant corpus of traditional Cornish (< 2 MByte) into memory. A new suite of more efficient programs (in FORTRAN) is being developed to process the data. Another improvement is the more judicious arrangement of the texts into blocks; e.g. the Ordinalia are separated into their three separate plays, because they were written by different scribes.

In the texts, words are spelled in variable orthographies; they may show initial mutation; nouns may show plural or singular suffixes, and verbs may be conjugated. In a dictionary, words are in a fixed orthography, and words with suffixes, if shown at all, are attached to a head-word. A key aspect of ***Nessa Tremen*** is the reduction of the words in traditional Cornish texts to a form suitable for publication in a dictionary. The process may be divided into three steps, from level 5 to level 2 in the following table:

LEVEL		*EXAMPLES*
1	ROOTS	
	determine the morphology	
2	HEAD-WORDS	***kemmeres*** 'to take'
	split into head-word and suffixes	
3	UNMUTATED TEXT	***kemmeras*** 'he/she/it took'
	remove initial mutations	
4	STANDARD TEXT	***gemmeras***
	put the word into ***Kernewek Kemmyn***	
5	ORIGINAL TEXT	*gemeras, gemerays, gomeras,* etc.

Although there are no plans at present to go above level 2 in ***Nessa Tremen***, it would be possible to go one stage further, and split words into their constituent morphemes (level 1).

Original text

Andrew Hawke kindly provided the author with computer-readable forms of most of the Cornish texts in their original spelling. After completing the corpus, reference numbers were added to all the lines (about 18000 of them). As an example, the following is the opening stanza of *Beunans Meriasek*, written by Radulphus Ton in 1504, with a literal translation into English appended.

BM.0001	me yw gylwys duk bryten /	I am called the Duke of Brittany,
BM.0002	ha seuys a goys ryel /	and risen from royal blood,
BM.0003	ha war an gwlascur cheften /	and a ruler in the kingdom,
BM.0004	nessa 3en myterne vhell /	second to the high king,
BM.0005	kyng conany /	King Conan;
BM.0006	aye lynneth purwyr y thof /	I am very truly of his lineage,
BM.0007	gwarthevyas war gvyls ha dof /	master over wild and tame,
BM.0008	doutis yn mysk arly3y /	feared among lords.

Standard text in *Kernewek Kemmyn*

The first problem is that the orthography is not fixed. This is not obvious from the eight lines shown here, but becomes so on reading a larger sample; e.g. for 'to the', Ton sometimes wrote *3en* (as here), and at other times *then*. Each word therefore needs to be referred in the first instance to the equivalent word in a standardized orthography. A double file was therefore made of each block of text, with the original text (level 5) on the left and the standard text in ***Kernewek Kemmyn*** (level 4) on the right. At the time of writing, not all blocks have yet been finished. The author is indebted to Keith Syed for making available versions of texts in ***Kernewek Kemmyn***.

BM.0001	me yw gylwys duk bryten /	My yw gelwys Dug Breten,
BM.0002	ha seuys a goys ryel /	ha sevys a woes ryal,
BM.0003	ha war an gwlascur cheften /	ha war an wlaskor chyften;
BM.0004	nessa 3en myterne vhell /	nessa dhe'n myghtern ughel
BM.0005	kyng conany /	King Konani:
BM.0006	aye lynneth pur~wyr y+th~of /	a'y linyeth pur wir yth ov,
BM.0007	gwarthevyas war gvyls ha dof /	gwarthevyas war wyls ha dov,
BM.0008	doutis yn mysk arly3y /	doutys yn mysk arlydhi.

Since the division of words is not always the same in both versions, the original text has been marked so as to correspond to the text in ***Kernewek Kemmyn***. The marker ~ is used to split a word, and the marker + is used to join two words. Thus the three words *purwyr y thof* is re-written as *pur~wyr y+th~of*, so that the division and number of words corresponds to the four words ***pur wir yth ov*** in the standard text. The other modification made at this stage is to remove all punctuation and most capitalization from the standard text, giving the following:

BM.0001	me yw gylwys duk bryten /	my yw gelwys dug Breten
BM.0002	ha seuys a goys ryel /	ha sevys a woes ryal
BM.0003	ha war an gwlascur cheften /	ha war an wlaskor chyften
BM.0004	nessa 3en myterne vhell /	nessa dhe'n myghtern ughel
BM.0005	kyng conany /	king Konani
BM.0006	aye lynneth pur~wyr y+th~of /	a'y linyeth pur wir yth ov
BM.0007	gwarthevyas war gvyls ha dof /	gwarthevyas war wyls ha dov
BM.0008	doutis yn mysk arly3y /	doutys yn mysk arlydhi

Direct lexical tagging

Direct lexical tagging would involve adding markers or tags to all words in the standard text which are not in the form of a head-word, to reduce them from level 4 to level 2. This process might produce a text like the following:

```
BM.0001  my bos>S13 gelwel>PP dug Breten
BM.0002  ha sevel>PP a{of} 2<goes ryal
BM.0003  ha war an 2<gwlaskor chyften
BM.0004  nes>CP dhe'n myghtern ughel
BM.0005  %king Konani
BM.0006  a'y{of his} linyeth pur wyr yth bos>S11
BM.0007  gwarthevyas war 2gwyls ha dov
BM.0008  doutya>PP yn mysk arloedh>PL
```

Here words with suffixes have been replaced by the appropriate head-word, followed by the symbol > and a code to denote the form of the suffix; words with initial mutation have been replaced by the root form, preceded by a number indication the type of mutation and the symbol <; and homographs have been distinguished by putting the English meaning after the word in curly brackets. In addition, the word *king* has been marked by the symbol %, denoting an unassimilated English word.

The more efficient alternative to direct lexical tagging

It is clear that to add lexical tags like these to the whole corpus would be a time-consuming task. Fortunately, there is no need to go down this road. The alternative, which involves much less labour, is to tag just the homographs. When these tags are included, the stanza becomes:

```
BM.0001  me yw gylwys duk bryten /          my yw gelwys dug breten
BM.0002  ha seuys a goys ryel /             ha sevys a7 woes ryal
BM.0003  ha war an gwlascur cheften /       ha war an wlaskor chyften
BM.0004  nessa 3en myterne vhell /          nessa dhe'n myghtern ughel
BM.0005  kyng conany /                      %king konani
BM.0006  aye lynneth pur~wyr y+th~of /      a'y2 linyeth pur wir yth ov
BM.0007  gwarthevyas war gvyls ha dof /     gwarthevyas war wyls ha dov
BM.0008  doutis yn mysk arly3y /            doutys yn mysk arlydhi
```

The only two words which have been tagged are ***a*** and ***a'y***. Cornish has several words spelled ***a***, here distinguished by numerical tags: ***a0*** 'O' (vocative, causing lenition); ***a1*** 'ah' (no mutation); ***a2*** (verbal particle); ***a4*** 'if' (causing provection); ***a6*** 'goes'; ***a7*** 'of'. The phrase ***a'y*** can mean 'of his' or 'of her', these being distinguished respectively by the numerical tags 2 and 3 (because these are the numbers referring to the mutations which they provoke).

In order to relate level 4 to level 2, it is necessary to set up a file which lists every different word at level 4. For this file unmutated and mutated forms (e.g. ***penn*** 'head', its lenited form ***benn*** and its spirantized form ***fenn)*** are listed separately. The file is known, rather prescriptively, as LAW.TXT (i.e. List of Allowable Words). The following small extract includes the word ***gelwys*** 'called', which appears in the stanza from *Beunans Meriasek.*

```
  2   gelmi          hm                    2kelmi
  7   gelwel         H     VN               gelwel          ***
  1   gelwes         d     73   OM.2774     gelwel
  2   gelwir         d     18               gelwel
  1   gelwis1        d     31   RD.0271     gelwel
  4   gelwis3        d     33   BM.4428     gelwel
 27   gelwys         H     AJ               gelwys          ***
  7   gemmer         dm    13               2kemmeres
  1   gemmera'       dmv   11   OM.1208     2kemmeres
  6   gemmeras       dm    33               2kemmeres
  1   gemmerav       dm    11   OM.1234     2kemmeres
  9   gemmeres       hm                    2kemmeres
```

The present state of the file is by no means definitive, since not all of the corpus has been processed. The extract does, however, give an indication of the methodology used. The columns represent, from left to right:

- the number of occurrences of the level 4 word (incomplete at present);
- the level 4 word, as it appears in the standard text, with tagging for homographs;
- a three-letter code denoting the status of the level 4 word:
  e.g. H = head-word, hm = mutated head-word, d = derivative, v = variant;
- a two-character grammatical code:
  e.g. VN = verbal noun, AJ = adjective, 11 = 1st person singular present indicative
- the line-number of *hapax legomena* (because the file is not yet complete, some may be false)
- the level 2 word, preceded where appropriate by a number denoting initial mutation:
  2 = lenition, 3 = spirantization, 4 = provection, 5 = mixed (as in text-books); here <g-> represents lenited /k-/ as well as /g-/.
- for the head-words, up to three stars showing in which edition of the dictionary the word will appear.

In this small extract, the following are noteworthy:

- The word ***gelwis*** is ambiguous, since it can mean 'I called' or 'he/she/it called'; it is therefore necessary to append a tag; here the numbers 1 and 3 are used respectively.
- Although the word ***gelwys*** 'called' is the past participle of ***gelwel*** 'to call', it is treated in the dictionary as a head-word and an adjective.
- The phrase *ny gemere* (level 5) found at *OM.1208* means 'I take not'; it is a variant of the lenited form of the 1st singular present indicative of ***kemmeres*** 'to take', the variation being occasioned by the loss of [-v].

There is no need to tag every occurrence of each suffixed word; instead, the referencing of such words to the relevant head-word is done in the file LAW.TXT. For instance, instead of separately tagging the six instances of ***gemmeras***, the referencing of this word to its head-form ***kemmeres*** is done once and once only. This idea, which is used in relational data-bases, saves work. Should it be nevertheless be desired to append lexical tags to each suffixed word, then software could be written to do this by machine.

Conclusion

A new computer-aided analysis of the Cornish texts, known as ***Nessa Tremen****, will use* ***Kernewek Kemmyn*** as a standard orthography for comparison. In relating individual words in the texts to the appropriate head-forms in a dictionary, it will not be necessary to add lexical tags to every mutated or suffixed word. Instead, the principle of relational data-bases is used, with purpose-written programs. This reduces the work-load considerably.

References

Dunbar, P. and George, K.J. (1997) ***Kernewek Kemmyn*** *– Cornish for the twenty-first century.* Cornish Language Board, Saltash.

Edwards, R. (1999) *Notennow Kernewek.* Notes on the Cornish texts published by *Kernewek dre Lyther,* Sutton Coldfield.

George, K.J. (1984) *The phonological history of Cornish.* Unpublished thesis for Doctorat du Troisième Cycle, University of Western Brittany, Brest.

George, K.J. (1986) *The pronunciation and spelling of Revived Cornish.* Cornish Language Board, Saltash.

George, K.J. (1988) 'The use of a mainframe computer to analyse the orthography of traditional Cornish'. *Proc. 1st North American Congress of Celtic Studies,* pp. 89-115.

George, K.J. (1993) *Gerlyver Kernewek Kemmyn - An Gerlyver Meur.* (a major Cornish - English dictionary) Cornish Language Board, Saltash.

Nance, R.M. (1929) *Cornish for all.* James Lanham, St Ives.