Lemmatisation of the Corpus of Cornish

Jon Mills
University of Luton

1. The Roots of the Cornish Language Revival

In this century the Cornish language has been revived so that today Cornish is spoken by many Cornish people as a second language. Moreover a few people currently living in Cornwall have been raised as bilingual from birth. Texts from the Middle Cornish (1200 to 1540 AD) and Modern Cornish (1540 to 1800 AD) periods form the basis upon which Cornish has been revived in the twentieth century. Dictionaries have made an important contribution to the pedagogical basis of this revival and several lexicographers have introduced standardised spelling systems for Cornish. The Williams' (1865) dictionary and Lewis' (1923) grammar form the basis of later twentieth century pedagogical dictionaries. Both Williams and Lewis are Welsh and have been led astray by analogy with the Welsh language on a number of issues. As a result, some subsequent dictionaries are not faithful to native Cornish practices. Analogy with Welsh and Breton has also led some revivalists to adopt Middle rather than Modern Cornish as the basis for standardising Cornish spelling and grammar. Furthermore Welsh and Breton have been used as sources for borrowing new words into twentieth century Cornish. Other writers, notably Jenner (1904) and Gendall (1997), have shown a preference for Modern Cornish as a pedagogical basis for the revival of Cornish. Jenner chose to take up the language where it had left off and disapproved of the inclusion of Welsh and Breton borrowings for which no authority exists in Cornish (Jenner 1904: xv). Jenner also adapted Modern Cornish orthography along loosely phonetic lines in order to create a consistent spelling system. Gendall (1997), on the other hand, has selected a single preferred spelling for each lexeme from those spellings that are attested in the historical literature. Today there are three different standardised spelling systems in common usage amongst revivalists, Unified Cornish (Nance 1938), Common Cornish (George 1993) and Modern Cornish (Gendall 1997).

In order to bring about a consensus on how to spell revived Cornish, reference needs to me made to the traditional material from the historical corpus. Historical Cornish is traditionally divided into three phases Old Cornish, Middle Cornish and Modern Cornish (Berresford Ellis 1974; George 1986: 8-10; Gendall 1991: ii). The Old Cornish phase lasted from 800 AD to 1200 AD. Few examples of Cornish survive from this period, the most important being the *Vocabularium Cornicum*, a Cornish Latin glossary. Appendix 1 lists the items of the Old Cornish phase. The Middle Cornish phase lasted from 1200 AD to 1540 AD. The principal texts of this period are all in verse and on a religious theme. Appendix 2 lists the items of the Middle Cornish phase. The Modern Cornish phase last from 1540 to 1800. This phase includes a far wider range of genres than the corpus of Old and Middle Cornish and includes Lhuyd's (1707) important account of the sound system and grammar. Appendix 3 lists the items of the Middle Cornish phase.

2. Lemmatisation

Lemmatisation may be defined as the creation of the base form corresponding to a given word form. This is usually achieved by transforming the word form. All the inflected forms of the unit are thus conventionally represented by the lemma so that, in English, for example, TAKE represents *take*, *takes*, *taking*, *taken*, and *took*. Lemmatisation may thus be considered a process of classification - that of words under their dictionary headword (Kipfer 1984: 166; Schnorr 1991: 2813). Existing dictionaries of Cornish suggest that approximately 9,000 lexemes can be obtained from the corpus. The lemmatisation of electronic text involves inserting a tag in the text for each occurrence of each lexeme in that text. Part-of-speech tagging separates most homographs though not those of identical word class. True lemmatisation of a text corpus involves inserting a tag that uniquely identifies each lexeme.

Inserting lexeme tags directly into text involves making decisions about segmentation. Whilst the *graphic word* may be defined as a string of one or more alphanumeric characters set off by a space or certain marks of punctuation on either side, this has its problems. For example, many compounds may be written either as separate words or hyphenated. One may wish to treat multi-word lexemes as units and orthographically united forms, such as enclitics, may be decomposed (Zampolli 1981: 242 *ff.*).

The lemmatisation system for the Corpus of Cornish needs to be able to handle a total of approximately 9,000 lexemes and relate these to their oblique forms so that it is possible to move from the forms to their lemmata and vice versa. The system must manage all the spelling and inflexional variants attested for the lexemes and provide direct access to the data contained in the corpus as well as the means for retrieval of selected information. Finally the system requires the capacity to expand the lexicon as new text is processed and new items are encountered.

Inconsistent orthography is a common problem for the corpus linguist especially in relation to older texts or those that represent diverging dialectal varieties. A multiplicity of spelling variants causes problems for the study of syntax or lexis (Rissanen 1994: 75; Markus 1994: 46; Milic 1994: 66). Within the corpus of Cornish orthographic variation is a major consideration. In their original form, the Cornish texts reflect the variety of orthographic styles, that were prevalent during the various chronological

episodes of the period they represent. The original spelling of the texts is not consistent, even normally within a single text. They are difficult to read in this form. For example, we find the following orthographic variants of the Cornish word for 'flesh': *chîc, cîg, cyc, gîc, gyc, gyke, kig, kìg, kîg, kyc, kych, kyek, kyg, kyk. kyke*. This leads to obvious difficulties when asking a computer to find a particular lemma for analysis, since a search has to be made, not only for all the inflected and mutated forms that the item can take, but also the many possible spellings of those. Such orthographic variation creates considerable difficulty for lemmatisation. Since many variants of the base form are attested, a way has to be found to provide lemmata that unify these.

It is frequently difficult to decide whether an item should be treated as one or more tokens. Nor have modern lexicographers completely solved problems of segmentation. For example, both Morton Nance (1938) and George (1993) treat **erbyn** (towards) as a lemma and write it as a single word. This item, however, is frequently found separated by an infixed possessive pronoun; for example

er y byn
towards him

Pascon agan Arluth stanza 29.

Cornish lexicographers sometimes disagree about segmentation. Morton Nance (1938) gives lemma status to **magata**, writing it as one word, unhyphenated. George (1993) does not give an entry for **magata**, but gives separate entries for **maga<as>** and **da<good>**.

Some linguists have suggested that morphological rules may be invoked to develop computer algorithms that achieve partial lemmatisation (Hellberg 1972; Martin, Al and van Sterkenburg 1983: 81; Garside, Leech & Sampson 1991: 152-55). However such an approach is not feasible for Cornish since morphological rules for Cornish cannot be established until the variant forms that comprise the lexicon are defined. Lemmatisation of the corpus must, therefore, precede the formulation of morphological rules.

As an alternative to the use of inflectional rules to automate lemmatisation, look-up dictionaries or databases are sometimes employed. By this method, raw text is matched against a machine dictionary which relates lemmata to their grammatical forms. The canonical form can be indexed to its grammatical words by means of a relational database written in the PROLOG programming language (Bien 1981). In the case of unambiguous forms, the lemma may be entered for the first occurrence and assigned automatically thereafter. Homographic forms may be supplied with alternative lemmata from which the lemmatiser selects manually.

Since unknown forms and lexical items may be expected, no computer software can possess all possible language forms. So the first text is lemmatised manually. The bank of forms thus generated, is applied to the second text. Those forms for which the form bank provides no lemmata are lemmatised manually, and the new forms are added to the bank. Thus the bank of graphical forms and lemmata continues to grow as lemmatisation proceeds. It is observed that different historical periods reflect different orthographies. So separate form banks for different historical periods are created (Busharia 1979: 136).

2.1. VOLTA: a lemmatisation algorithm

It is not possible to compile an inventory of lexeme tags in advance of the lemmatisation process because a complete inventory is not established until the process is complete. So the lemmalist from George's 1993 dictionary was used as an initial guide to segmentation and as a provisional set of lexeme tags. Additional lexeme tags had to be created for lexemes that are not listed in the George dictionary.

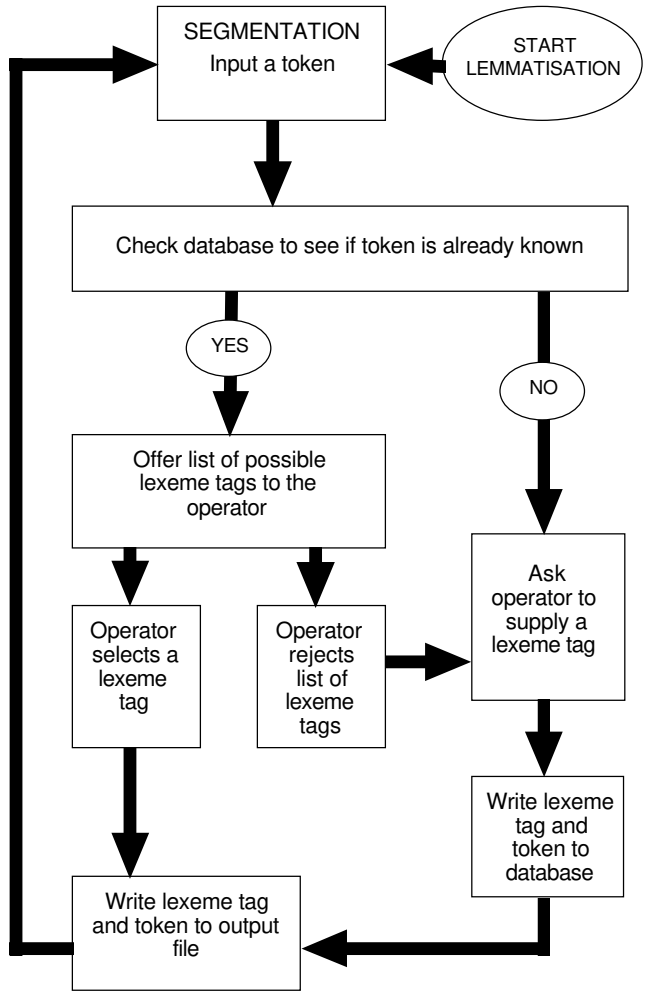

Figure 1

Computer assisted tagging of the corpus was achieved with the aid of a specially written program called VOLTA (Vertical Output Lemma Tagging Aid). Figure 1 illustrates the algorithm. Characters are input from the raw text file until a whole word token is captured. Segmentation is thus decided by the human operator. The operator is then asked to supply a lexeme tag for this token. Figure 2 shows the computer screen with lemmatisation in progress. The token and its tag are written to an output file. At the same time they are added to an object oriented knowledge base of lexeme tags and variant forms. The process is repeated. Next time the same word type is encountered VOLTA prompts the operator asking if the tag that was previously assigned to the type is the one required. If the operator confirms the choice it is automatically written to the output file. Alternatively the operator can enter a different tag. Outer selection is thus a continuous process in which the database grows as lemmatisation proceeds.

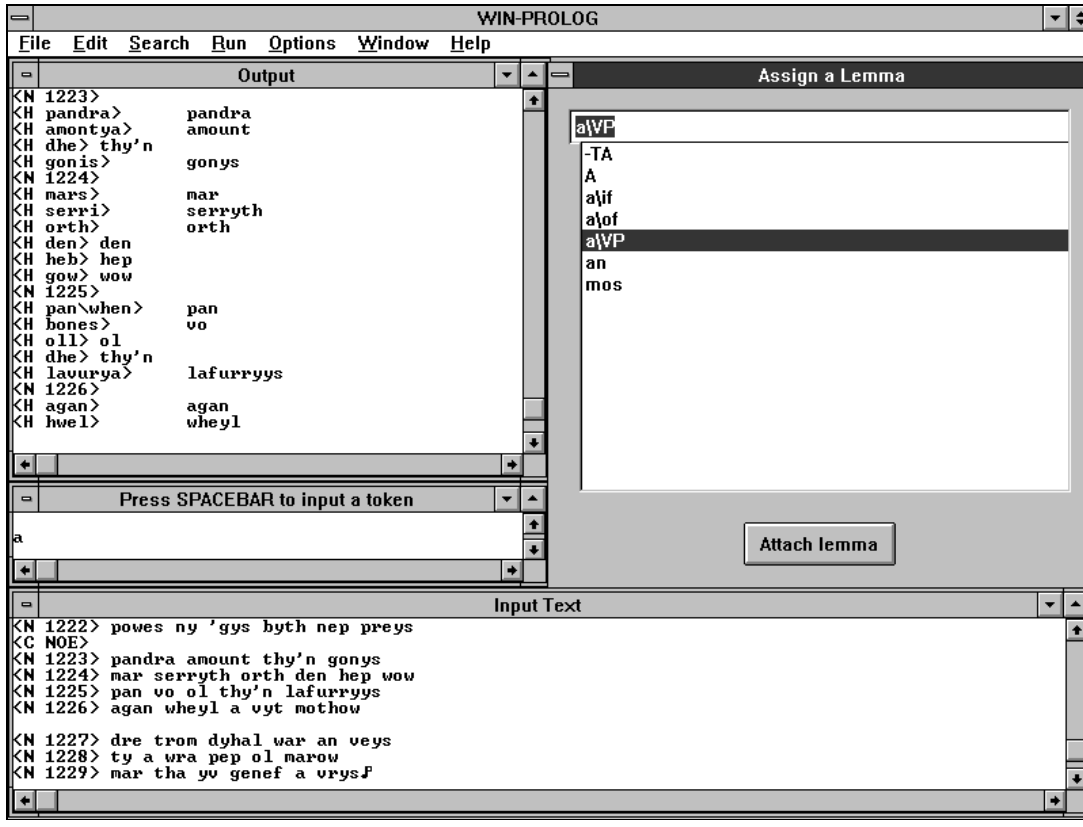

Figure 2

The database consists of the lexeme tag and its type entered in the form, l(*Lexeme-tag*, *Type*). Figure 3 shows the database generated from *William Allen's Cornish Rhyme* a short text of 23 running words in length.

```
l(kynsa,           kensa).
l(blydhen,         blethan).
l(byrla,           byrla).
l(ha,              'a'''').
l(baya,            baye).
l(nessa,           nessa).
l('lull-ha-lay',   'lull a'' laye').
l(tressa,          tridgya).
l(hedhes,          hanna).
l(dri,             dr).
l(omma,            ubba).
l(peswara,         peswarra).
l(blydhen,         bletha).
l(molleth,         mola).
l('Dyw',           'Dew').
l('war\on',        war).
l(ev,              ef).
l(gul,             weeg).
l(dri,             dry).
l(hi,              hy).
l(omma,            uppa).
```

Figure 3

The lemmatised output file is arranged vertically with one word token per line. Comments are placed in square brackets (see Figure 4). There are two reference fields entered as COCOA references in angled brackets. References preceded by "N" refer to line numbers in the original text. The lexeme tag is entered as a COCOA reference, <H *Lexeme*> followed by a tab and the Token.

```
["William Allen's Cornish Rhyme" OC IV.9]
<N 1><H kynsa>          Kensa
<H blydhen>             blethan
<H ,>                   ,
<H byrla>               byrla
<H ha>                  a'
<H baya>                baye
<H :>                   :
<N 2><H nessa>          Nessa
<H blydhen>             blethan
<H ,>                   ,
<H lull-ha-lay>         lull a' laye
<H ;>                   ;
<N 3><H tressa>         Tridgya
<H blydhen>             blethan
<H ,>                   ,
<H hedhes>              hanna
<H dri>                 dr
<H omma>                +ubba
<H ,>                   ,
<N 4><H peswara>        Peswarra
<H blydhen>             bletha
<H ,>                   ,
<H molleth>             mola
<H Dyw>                 Dew
<H war\on>              war
<H ev>                  ef
<H gul>                 weeg
<H dri>                 dry
<H hi>                  hy
<H omma>                uppa
<H .>                   .
```

Figure 4

Segmentation is indicated by placing each token on a new line. Tokens may include blank spaces. In order to show that, in the original text, an item is attached to the preceding word without an intervening white space the "+" sign is entered.

Each text or set of texts is lemmatised separately with its own lemmatisation database of lexeme tags and word types. This is for two reasons. Firstly, since there is frequently considerable orthographic variation between texts, shared databases between texts are not always particularly efficient. Secondly, a dictionary of word forms for each text can be separately generated by means of a specially written program. Figure 5 shows such a dictionary generated from *William Allen's Cornish Rhyme*, in which entries are arranged alphabetically by lexeme tag.

```
baya:         [baye]
blydhen:      [bletha, blethan]
byrla:        [byrla]
dri:          [dr, dry]
Dyw:          [Dew]
ev:           [ef]
gul:          [weeg]
ha:           [a']
hedhes:       [hanna]
hi:           [hy]
kynsa:        [kensa]
lull-ha-lay:  [lull a' laye]
molleth:      [mola]
nessa:        [nessa]
omma:         [ubba, uppa]
peswara:      [peswarra]
tressa:      tridgya]
war\on:       [war]
```

Figure 5

Finally contexts may be accessed via *Micro-OCP* concordancing software. Figure 6 shows a complete lemmatised KWIC concordance to *William Allen's Cornish Rhyme*.

```
baya
   Kensa blethan ,byrla a' baye :Nessa blethan ,lull a' lay
blydhen
                      Kensa blethan ,byrla a' baye :Nessa bl
    ,lull a' laye ;Tridgya blethan ,hanna dr +ubba ,Peswarr
   anna dr +ubba ,Peswarra bletha ,mola Dew war ef weeg dry
   n ,byrla a' baye :Nessa blethan ,lull a' laye ;Tridgya b
byrla
            Kensa blethan ,byrla a' baye :Nessa blethan ,lu
dri
   ;Tridgya blethan ,hanna dr +ubba ,Peswarra bletha ,mola
   a ,mola Dew war ef weeg dry hy uppa
Dyw
    ,Peswarra bletha ,mola Dew war ef weeg dry hy uppa
ev
   ra bletha ,mola Dew war ef weeg dry hy uppa
gul
   bletha ,mola Dew war ef weeg dry hy uppa
ha
      Kensa blethan ,byrla a' baye :Nessa blethan ,lull a'
hedhes
    laye ;Tridgya blethan ,hanna dr +ubba ,Peswarra bletha
hi
   ola Dew war ef weeg dry hy uppa
kynsa
                            Kensa blethan ,byrla a' baye :Ne
lull-ha-lay
   a' baye :Nessa blethan ,lull a' laye ;Tridgya blethan ,h
molleth
   +ubba ,Peswarra bletha ,mola Dew war ef weeg dry hy uppa
nessa
   blethan ,byrla a' baye :Nessa blethan ,lull a' laye ;Tri
omma
    Dew war ef weeg dry hy uppa
   idgya blethan ,hanna dr +ubba ,Peswarra bletha ,mola Dew
peswara
   lethan ,hanna dr +ubba ,Peswarra bletha ,mola Dew war ef
tressa
    blethan ,lull a' laye ;Tridgya blethan ,hanna dr +ubba
war\on
   swarra bletha ,mola Dew war ef weeg dry hy uppa
```

Figure 6

2.2. Improving VOLTA

Tokenising the text manually proved slow and bothersome. It was, therefore, decided to develop a way of semi-automating tokenisation that could be integrated into VOLTA.

Sentence word tokenisation is the process of converting a sentence into a string of words. We tend to take for granted the unit that we loosely call a word. In languages such as English, words are delimited by spaces and punctuation marks. The task of tokenisation may, consequently, seem trivial. The medieval manuscripts that are the sources for the Corpus of Cornish, however, do not always have explicit token delimiters. Such orthographic practice makes tokenisation a serious problem.

The token-type distinction originates with the philosopher Charles Peirce. However Peirce discriminates between tokens, types and tones. These three emerge as manifestations of three modes of reality: existential reality, the reality of law, and the reality of qualities. Tokens equate with existential reality. In other words, a token belongs to the existential world. When we say that a sign is a token, we are indicating what is absolutely unique in its occurrence, its place in time and space (Winder 1996). With regard to a text corpus, tokens are, thus, simply text positions.

Types equate with the reality of law. A type is a sign that represents the law-like generality of a class. Unlike tokens, we cannot point to a type anymore than we can point to the law of gravity. Although types are real, they do not belong to the existential world in which pointing is possible (Winder 1996). With regard to a text, one can say that it has a vocabulary of X number of word types.

Tones equate with the reality of qualities. In any given investigation, there are certain perceptual units that cannot or will not be analysed. These are qualities. Although we may recognise the general form of a scrawl, we may not distinguish individual letters. We cognise some quality whether we recognise the form or not. In other words, whatever interpretation we may finally bring to something, our first impression has a value which is distinct from time and space and distinct from law. That value is tone. With regard to text, one has an awareness of more the one actually uses for reading purposes. For example a letter may be smeared and the spacing of the text may vary. Such qualities may be perceived but the reader can choose to ignore them. The tones of a text are, thus, defined as those qualities which one does indeed wish to consider as fundamental yet unanalysable in a given analysis. For example, alphanumeric characters are usually considered unanalysable. In other words, the reader does not analyse them into bars and curves, or distinguish them according to pitch in proportional spacing (Winder 1996).

Figure 7 shows an extract from the 16th century Cornish miracle play, *Beunans Meriasek* (Ton 1504: 19).

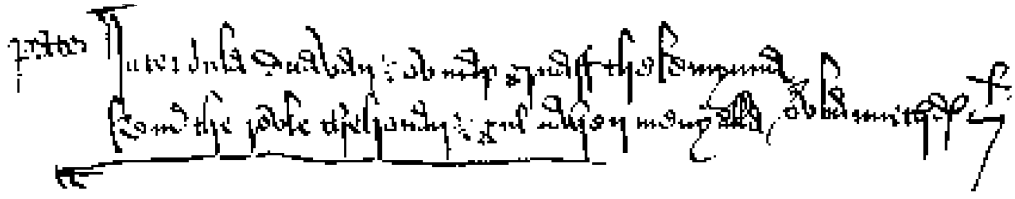

Figure 7

Below is a transcription and translation of the above extract.

Inter dula du avan	*Between the hands of God above*
ov map gruaff the kemynna	*My son, I do commend thee*
kemmer the roule the honan	*Take thine own rule:*
gul nahen me ny alla	*Do aught else I cannot*
ov banneth dis	*My blessing to thee*

From the extract it can be seen that the two lines of handwriting in the original actually represent five lines of verse. That it is indeed verse is evident from the rhyming of alternate lines and the metre of seven syllables per line. The transcription shows three instances of the word ‘the’. These occurrences represent three word tokens but only one word type. Figure 8 shows the tone of the first occurrence of ‘the’.

Figure 8

There are essentially two approaches to word tokenisation, character based tokenisation and lexicon based tokenisation. Character based tokenisation assumes that there are certain characters, such as the letters of the alphabet, that occur within words. It is similarly assumed that certain other characters, such as spaces and punctuation, occur between words. There can be, however, certain problems with these assumptions. Whilst character based tokenisation works fairly well for some languages such as English, the medieval Cornish texts do not always have explicit word delimiters. In the extract from *Beunans Meriasek* in Figure 7, it can be seen that spaces between words are frequently omitted or not clear.

Lexicon based tokenisation is allied to Peirce’s concepts of token, type and tone. The very notion of type implies a lexicon. The lexicon is, thus, derived from the text. Whilst, conversely and simultaneously, knowledge of the lexicon enables interpretation of the text. Figure 9 illustrates an algorithm for lexicon based tokenisation.

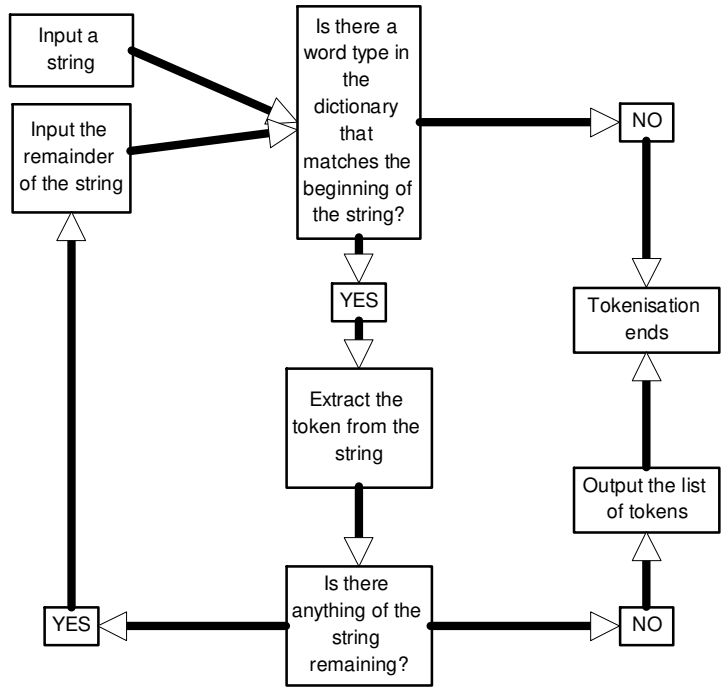

Figure 9

It is necessary to represent a text that has been tokenised in such a way that both individual tokens and individual types can be identified. Lager (1996 : 34) describes how segments and points form a basis for tokenisation . A critical point in a text is that which delimits two adjacent segments. Conversely a critical segment of text is that which is delimited by two points in the text. Segments are, thus, located in time and space. Segments can be observed, pointed at and given unique names. Segments and strings are not the same. Segments are instances of strings. Two segments may be instances of the same string. Segments are, thus, tokens, whereas strings are types. Figure 10shows how critical tokenisation can be applied to the first line of the fragment of the Cornish miracle play shown in Figure 7.

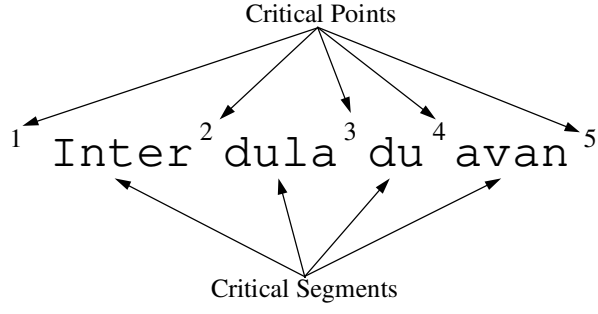

Figure 10

Figure 11 shows how the principles of critical tokenisation may be applied to the fragment of the Cornish miracle play shown in Figure 7 in order to create a Prolog database. The text is arranged vertically with each token on a separate line (see Figure 11). Each token is represented by a three place predicate. The first and second arguments are the critical points which define the segment covered by the token. The third argument is the word type of which the token is an instance.

```
token( 1,  2, inter).
token( 2,  3, dula).
token( 3,  4, du).
token( 4,  5, avan).
token( 5,  6, ov).
token( 6,  7, map).
token( 7,  8, gruaff).
token( 8,  9, the).
token( 9, 10, kemynna).
token(10, 11, kemmer).
token(12, 13, the).
token(13, 14, roule).
token(14, 15, the).
token(15, 16, honan).
token(16, 17, gul).
token(17, 18, nahen).
token(18, 19, me).
token(19, 20, ny).
token(21, 22, alla).
token(23, 24, ov).
token(24, 25, banneth).
token(25, 26, dis).
```

Figure 11

With the text in the form of a database it is possible to conduct various types of search. For example, if one wants to know what word type is found between critical points 2 and 3, one sets the goal,

```
?- token(2,3,T).
```

The system returns,

```
T = dula
```

Conversely, if one wants to know in what text positions the word type 'dula' is found, one sets the goal,

```
?- token(X,Y,dula).
```

The system returns,

```
X = 2
Y = 3
```

If one wants the phrase found between critical points 1 and 5, one sets the goal,

```
?- get_segment(1,5,Segment).
```

The system returns,

```
Segment = [inter,dula,du,avan] .
```

If one wants an alphabetical list of all the word types attested in the text, one sets the goal,

```
?-setof(T,X^Y^token(X,Y,T),List).
```

The system returns,

```
List = [avan,du,dula,gruaff,inter,kemmer,kemynna,map,ov,the]
```

The process is, thus, seen to be founded on a type of quotation, which we call an 'attestation'. In this manner, routines can be written in Prolog to produce frequency wordlists, concordances, lists of collocations and other types of data output.

3. The Future

As this millennium draws to a close, the Corpus of Cornish could be extended to include a subcorpus of twentieth century Cornish. This subcorpus should include a variety of different spelling systems, styles and genres. The completed lemmatised Corpus of Cornish will provide data that engenders the study of Cornish syntax, morphology, lexicology and pragmatics. Finally movement towards a consensus on a spelling system for Cornish in the twenty first century may be encouraged and facilitated by a better understanding of Cornish linguistic heritage

Appendix 1

Smaragdus's Commentary on Donatus, Cornish glosses	End of 9th century	Paris Bibliotheque Nat. MS.Lat. 13029
Oxoniensis Posterior, 3 Cornish glosses	10th century	Bodleian 572
Manumissions in the Bodmin Gospels	10th century	British Library Add. 9381
Vocabularium Cornicum	circa 1100	British Library Cott Vesp A xiv
Joannis Cornubiensis (John of Cornwall) "Prophetia Merlini"	12th century	Vatican Cod. Ottobonianus Lat. 1474

Appendix 2

The Glasney Cartulary	13th century quoted in a 15th century manuscript	MS Cornwall County Records Office Dd R(S)59
The Charter Endorsement	late 14th century	British Library Add Cart 19491
Pascon agan Arluth	15th century	British Library Harl N. 1782
The Ordinalia	circa 1500	Bodleian 791
Beunans Meriasek	1504	National Library of Wales. MS. Peniarth 105
“Black Book of Merthen”	1506-1536	Royal Institution of Cornwall

Appendix 3

Fyrst Boke of the Introduction of Knowledge by Andrew Boorde	1547.	British Library C.71.b.29
Star Chambers		
Image of Idleness	circa 1550	
John Tregear Homilies	1560	British Library Add 46397
Exeter Consistory Court Depositions	1569-1572	Henderson MSS Truro Museum vol. X : 176
The Survey of Cornwall by Richard Carew	1602	British Library 291.e.3
Gwreans an Bys	1611	Bodleian 219
Richard Brome The Northern Lass	1632	Printed by Aug. Mathevves
Diary of Richard Symonds	1644	British Library Add 17062
MS of Keigwin's including “A protestation of the bishops in Britain to Augustine the monk, the pope's legate in the year 600 after Christ.”, “First chapter of Genesis”,		Bodleian Gough Cornwall 3-4

Gwavas collection: Wella Kerew alias William Rowe's translations of: (a) Genesis III, (b) St. Matthew II, (c) St. Matthew IV.	late 17th century	British Library Add 28,554
Scawen “Antiquities Cornu-Britannick or Observations on an Ancient manuscript Entitled Passio Christi”	Written circa 1680 published London 1777	
Various		Bonaparte MSS Bilbao Library
William Hals “An Lhadymer ay Kernou”		National Library of Wales 1662
William Hals' history. Words used by Revd John Jackman in administering the sacrament in the 17th century.	Early 18th century	British Library Add 29762
Penzance Manuscript		Courtney Library
Thomas Tonkin’s Manuscript B	Early 18th century	Royal Institution of Cornwall
Thomas Tonkin’s Manuscript H	Early 18th century	Royal Institution of Cornwall
Edward Lhuyd Archaeologia Britannica	1707	Published Oxford: The Theatre.
William Gwavas’ Notebook	1710	Royal Institution of Cornwall
Jottings by William Gwavas on reverse of a legal document	1732	Royal Institution of Cornwall
William Gwavas’ copy of John Boson’s Pilchard -Curing Rhyme		Royal Institution of Cornwall
William Gwavas’ copy of John Boson’s Ten Commandments, Lord’s Prayer and Creed		Royal Institution of Cornwall
Borlase, W. “Mems. of Cornish Tongue” Copies of MSS. by Lhuyd, Gwavas, Tonkin, Ustick, Scawen and Boson in Borlase's own hand.	1748	Unpubl. MSS. in the Royal Institute of Cornwall.
Manuscript of Nicholas Boson’s “Nebbaz Gerriau dro tho Carnoack”, Henry Ustick’s Hand	1750	Royal Institution of Cornwall
The Scawen Manuscript (in Tonkin’s Hand)		British Library Add 33420
Scawen, W. Oberservations on a Manuscript Entitled Passio Christi..	1777	British Library 628.k.17
Lhuyd, E. Geirlyfr Kyrnweig		NLW Llanstephan 84
Lhuyd’s phonetically spelled transcript of James Jenkins’ verses		Bodleian 10712

MS Belonging To Lhuyd, Extracts From "The Dutchess Of Cornwall's Progresse To See Ye Land's End And To Visit Ye Mount" Nicholas Boson.		Bodleian 10,714
William Bodinar's Letter	(c. 1778)	Original in the possession of the Society of Antiquaries
John Whitaker “Ruan Lanihorne”		Royal Institution of Cornwall

Works Cited

Berresford Ellis, P. (1974) *The Cornish Language and its Literature* London: Routledge and Kegan Paul.

Bien, J. (1981) "Toward Computerised Dictionaries for Inflexional Languages: Internal Design Philosophy" in Zampolli and Cappelli eds. (1983 : 77-82).

Busharia, Z. (1979) "Computerized Lemmatization of Non-vocalised Hebrew Texts" Proceedings of the International Conference on Literary and Linguistic Computing Tel Aviv: Tel Aviv University, Katz Research Institute 133-40.

Garside, R., G. Leech & G. Sampson eds. (1991) *The Computational Analysis of English: a Corpus Based Approach* Oxford: OUP.

Gendall, R.R.M (1991a) *A Students' Grammar of Modern Cornish - Part 1, English Cornish* Menheniot: The Cornish Language Council.

Gendall, R.R.M. (1991b) *A Students' Dictionary of Modern Cornish: Part 1 English-Cornish* 3rd ed. Menheniot: The Cornish Language Council.

Gendall, R.R.M. (1997) *A Practical Dictionary of Modern Cornish: Part One Cornish - English* Menheniot: Teere ha Tavaz.

George, Ken (1986) *The Pronunciation and Spelling of Revived Cornish* n.p.: The Cornish Language Board.

George, Ken (1993) *Gerlyver Kernewek Kemmyn: An Gerlyver Meur* n.p.: The Cornish Language Board.

Hausmann, F. (1991) *An International Encyclopedia of Lexicography* Berlin: Walter de Gruyter.

Hellberg, S. (1972) "Computerized Lemmatization without the Use of a Dictionary: a Case Study from Swedish Lexicology" *Computers and the Humanities* VI 209-12.

Jenner, Henry (1904) *Handbook of the Cornish Language* London: David Nutt.

Kipfer, Barbara Ann (1984) *Workbook on Lexicography: A Course for Dictionary Users with a Glossary of English Lexicographical Terms* Exeter Linguistic Studies Vol. 8. Exeter: University of Exeter Press.

Kytö, M., M. Rissanen & S. Wright eds. (1994) *Corpora across the Centuries* Proceedings of the First International Colloquium on English Diachronic Corpora. Language and Computers: Studies in Practical Linguistics. Amsterdam: Rodopi.

Lager, Torbjörn (1995) *A Logical Approach to Computational Corpus Linguistics* Gothenburg Monographs in Linguistics 14. Gothenburg: Department of Linguistics, Göteborg University.

Lewis, Henry (1923) *Llawlyfr Cernyweg Canol* Wrecsam: Hughes.

Lhuyd, E. (1707) *Archaeologia Britannica: Vol. I Glossography* Oxford: The Theatre.

Markus, M. (1994) *The Concept of ICAMET (Innsbruck Computer Archive of Middle English Texts)* in M. Kytö, M. Rissanen & S. Wright eds. (1994).

Martin, W., B. Al & P. van Sterkenburg (1983) "On the Processing of a Text Corpus: from Textual Data to Lexicographical Information" in Hartmann ed. (1983).

Milic≅, L. (1994) *Is a Grapheme a Lexeme? And other Problems of Definition in the Century of Prose Corpus* in M. Kytö, M. Rissanen & S. Wright eds. (1994).

Morton Nance, R. (1938) *New Cornish - English Dictionary* n.p.: Federation of Old Cornwall Societies.

Rissanen, M. (1994) “The Helsinki Corpus of English Texts” in M. Kyto", M. Rissanen and S. Wright. (1994) eds..

Schnorr, V. (1991) “Problems of Lemmatisation in the Bilingual Dictionary” in F.Hausmann (1991).

Ton, Rad (1504) *Beunans Meriasek* Ms. Peniarth 105. National Library of Wales.

Williams, R. (1865) *Lexicon Cornu-Britannicum - Gerlyvr Cernewec* London: Trubner.

Winder, W. (1996) "Reading the Text's Mind: Lemmatisation and Interpretation from a Peircean Perspective" Available http://www.chass.utoronto.ca:8080/epc/chwp/winder/.

Zampolli, A. & A. Cappelli eds. (1983) *Linguistica Computazionale III: The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries* Proceedings of the European Science Foundation Workshop, Pisa, 1981.

Zampolli, A. (1981) "Lexicological and Lexicographical Activities at the Istituto di Linguistica Computazionale" in Zampolli and Cappelli eds. (1983: 237-78).