Computer-assisted Lemmatisation of a Cornish Text Corpus for Lexicographical Purposes

Jon Mills

PhD Thesis

University of Exeter

2002

Abstract

This project sets out to discover and develop techniques for the lemmatisation of a historical corpus of the Cornish language in order that a lemmatised dictionary macrostructure can be generated from the corpus. The system should be capable of uniquely identifying every lexical item that is attested in the corpus. A survey of published and unpublished Cornish dictionaries, glossaries and lexicographical notes was carried out. A corpus was compiled incorporating specially prepared new critical editions. An investigation into the history of Cornish lemmatisation was undertaken. A systemic description of Cornish inflection was written. Three methods of corpus lemmatisation were trialed. Findings were as follows. Lexicographical history shapes current Cornish lexicographical practice. Lexicon based tokenisation has advantages over character based tokenisation. System networks provide the means to generate base forms from attested word types. Grammatical difference is the most reliable way of disambiguating homographs. A lemma that contains three fields, the canonical form, the part-of-speech and a semantic field label, provides of a unique code for every lexeme attested in the corpus. Programs which involve human interaction during the lemmatisation process allow bootstrapping of the lemmatisation database. Computerised morphological processing may be used at least to partially create the lemmatisation database. Disambiguation of at least some of the most common homographs may be automated by the use of computer programs.

Table of Contents

Table of Figures

1 Introduction

Cornwall is situated in the south-west peninsula of the island of Britain in the European Archipelago. Cornish, the language of Cornwall, is a Brythonic Celtic Language. It is usually thought that Cornish died out at the end of the eighteenth century (Berresford Ellis 1974; Pool 1982). Today, however, Cornish is undergoing revival and is spoken by several hundred people in Cornwall (EKOS & SGRÙD 2000: 45). The corpus of historical Cornish prior to the revival consists of texts from the Middle Cornish (1200 to 1575 AD) and Modern Cornish (1575 to 1800 AD) periods. It is this corpus with which this project is concerned.

A variety of reference sources provide information about the Cornish lexicon over a period of approximately a thousand years. Glosses in the margins of Latin manuscripts give Cornish equivalents for items in the text. Glossaries provide lists of items with their equivalents. The notes and essays of philologists explore an assortment of data concerning lexical items. Published and unpublished dictionaries give more comprehensive accounts of the Cornish lexicon. Cornish lexicography has passed through three phases. During the first phase, which includes the early glosses and the *Vocabularium Cornicum* (VC), the target language is Latin and the dictionary user's first language Cornish. The second phase begins in the mid 17th century and is purely descriptive. In other words the lexicographer is simply recording data about the Cornish lexicon. Meaning is dealt with by providing English translation equivalents. This overlaps with the third phase, in which

reconstruction is attempted by the lexicographer. Lhuyd (AB), in 1707, is the first to fill in gaps in the lexicon by borrowing from Welsh. He is followed in 1769 by Borlase (VCBL) and in the twentieth century by Morton Nance (NCED, ECD2, ECD3, CED). In the 20th century, several attempts have been made to standardise spellings to meet the demands of Cornish language revivalists (Morton Nance 1929; George 1986; PDMC).

The general methodology of lexicography has been described in a number of works (Partridge 1963; Zgusta 1971; Hartmann ed. 1983; Landau 1989; Hausmann et al. 1989-1991; Svensén 1993; Newell 1995). These methodologies are mainly oriented towards the major languages of the world, especially English. Cornish, as with all languages, has its own lexicographical idiosyncrasies. It is usual for a text corpus to serve as a basis for constructing a dictionary. In recent years, computer technology has come to play an increasingly important role with regard to the computational storage of the lexicon (see Ooi 1998; Walker, Zampolli & Calzolari eds. 1994; Atkins & Zampolli eds. 1994; Guo ed. 1995) and corpus based lexical modelling of language (see Sinclair 1991; Flowerdew & Tong eds. 1994; Lager 1995; McEnery & Wilson 1996; Thomas & Short eds. 1996; Kennedy 1998). Thus, nowadays, it is common for the corpus to consist of a number of computer files.

Central to lexicography is the notion of lemmatisation. Lemmatisation is sometimes defined as the “creation of the base form corresponding to a given word form, usually achieved by transforming the word form” (Schnorr 1991: 2813). All the inflected forms of the unit are thus conventionally

represented by the lemma: *umbrella* for *umbrella* and *umbrellas*, *take* for *take*, *takes*, *taking*, *taken*, and *took*, or *go*, for, *go*, *goes*, *going*, *gone*, and *went*. In this manner inflected forms are normally all treated together in the same entry, under the same entry form (Béjoint 1994: 192). Lemmatisation may thus be considered a process of "classification - that of words under their dictionary headword" (Kipfer 1984: 166).

Dictionary word lists, however, are not always restricted to base forms. Oblique forms are also included in the word lists of some dictionaries; Williams' *Lexicon Cornu-Britannicum* (LCB) is a case in point. If the lemma is seen as that part of the entry which determines the position of the entry in the word list (Ilson 1988; Hausmann & Wiegand 1989-1991; Osselton 1995; Hartmann & James 1998), then lemmatisation may be redefined as the process which determines the ordering of the word list in the dictionary macrostructure (Schnorr 1991; Botha 1992; Lorentzen 1996). The reduction of a paradigm of variant forms to its base form, then, is one form of lemmatisation, which I shall refer to as base form lemmatisation.

According to Muller (1977: 6), the laws of lemmatisation are entirely conventional. However the conventions have occasionally been challenged. Matoré (1968: 191) considers that the dictionary presents an arbitrary picture of the language and points to lexicographical practices which, whilst sanctioned by tradition, might be considered debatable. Why, for example, should nouns be presented without an article in the masculine singular form. And, why should verbs be presented in the infinitive, even though that form may be relatively little used. Béjoint (1994: 192), nevertheless, points out

that, if it is accepted that lemmatisation rules are only arbitrary, but convenient, conventions, there is no need for change.

1.1 Nature and scope of problem

For lexicographical or lexicological purposes one may wish to consult a concordance of a given lexeme. Normally a concordance does not arrange its entries according to their lemmata. As a result, word-types that belong to the same lexeme are distributed throughout the concordance and do not necessarily appear adjacent to one another. In the case of a language such as Cornish, which not only displays considerable inflectional variation but also undergoes mutation of initial consonants, the problem is quite severe. In the case of the historical corpus of Cornish, in which spelling is capricious, this problem is compounded.

The base form lemmatisation of an electronic text corpus involves inserting a tag in the text for each occurrence of each lexeme in that text. A well contrived system of corpus lemmatisation is essential in order for the Corpus of Cornish to be accessible to techniques of electronic text analysis and retrieval. Whilst the principles of lemmatisation in a dictionary are relatively well understood, the methodology of corpus lemmatisation has its own considerations to be taken into account.

The aim of this project, then, is to discover and develop a technique for lemmatisation of a historical corpus of the Cornish language. The research question, then, is what methods and techniques can be brought to bear on a historical corpus of Cornish to generate a lemmatised dictionary

macrostructure? The system should be able to cope with every lexical item contained in the historical corpus of Cornish. Lemmata should provide a unique code for every lexeme attested in the corpus.

1.2 Method of investigation

A survey of existing Cornish dictionaries, glosses and lexicographical notes was undertaken in order, firstly, to determine what is already known about the Cornish lexicon and, secondly, to identify lexicographical tradition. A corpus was compiled. Where possible, digitised images of the manuscripts were obtained. Published critical editions of the manuscripts and editions in normalised spelling were obtained. From these my own digital editions were prepared and it is these which comprise the electronic corpus. A method for tokenising the corpus was devised. Programs were written to perform lexicon based tokenisation and character based tokenisation. The corpus was tokenised using a combination of these two methods. The historical development of the lemma in Cornish lexicography was traced and an analysis of lemmatisation in Cornish dictionaries was undertaken. The principles of alphabetisation of the word list are discussed. An analysis of lexical variation of form is undertaken in order to show the formal relationship between the canonical form chosen as a head word and all its variant forms that are attested in the corpus. The description of the inflection system of Cornish is new. The methodology for disambiguating homographs is discussed. An important criterion for distinguishing between homographs is their part-of-speech. It was necessary, therefore, to determine what criteria might be employed for the identification

of part-of-speech in the corpus of Cornish.

Approaches to corpus lemmatisation are discussed and three methods of corpus lemmatisation were formulated and then trialed. The first method involves looking up the form of tokens in a dictionary to determine their base form. Special software was developed for this purpose. The second method involves aligning the corpus in its original spelling with a version in a normalised spelling. The normalised tokens are looked up in a dictionary to determine their base form as for the first approach. The third method involves aligning the corpus in its original spelling with a version in a normalised spelling. The base form is then generated from a normalised form by the application of morphological rules. Using a combination of these three methods, a lemmatised concordance of the entire corpus was produced.

Underlying the methodology described in this thesis is the notion of System which is borrowed from Systemic Linguistics. The concept of System within linguistics originates with Firth (1957) and was later developed by Halliday (1956, 1961). In this project, system networks are used to represent and encode the morphology and syntax relating to lexical items. It is from this morphological-syntactic system network that the base form is generated. System networks are also used to represent options within the lemmatisation process. The resulting method might be termed Generative Systemic Lemmatisation.

A system begins with a point of origin and may operate with hierarchical or simultaneous entry conditions. The **[** symbol represents logical disjunction, the

Boolean operator EITHER/OR. Figure 1 illustrates a hierarchical system in which either [a] or [b] is chosen. And if [b] is chosen, then one continues by choosing either [c] or [d].

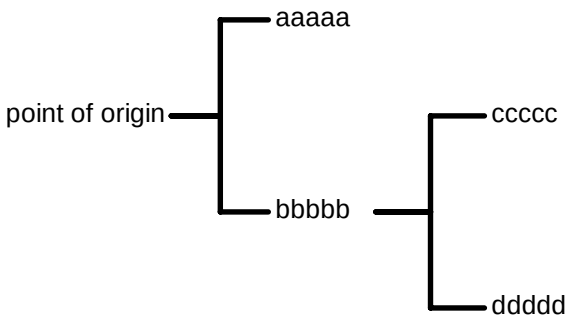


Figure 1 Hierarchical system

The **{** symbol represents logical conjunction, the Boolean operator AND. Figure 2 illustrates a simultaneous system in which both [a] and [b] are chosen. And [b] entails the further choice of both [c] and [d].

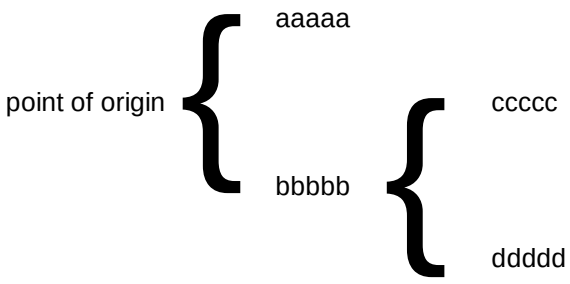


Figure 2 Simultaneous system

The entry conditions to a point in a system may be simple, compound or

disjunctive. A simple system, such as the one illustrated in Figure 3, requires that only feature [a] be chosen before a further choice of [c] or [d] is required.

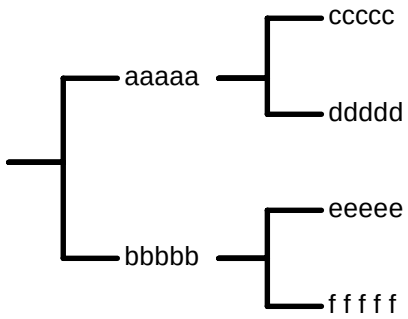


Figure 3 Simple system

A compound system, such as the one illustrated in Figure 4, requires that both features [b] and [c] apply before a further choice between [e] and [f] is made.

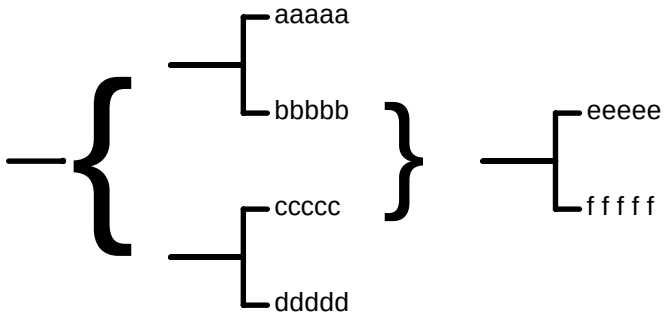


Figure 4 Compound system

A disjunctive system, such as the one illustrated in Figure 5, requires that either [b] or [c] is chosen before a further choice between [e] and [f] is made.

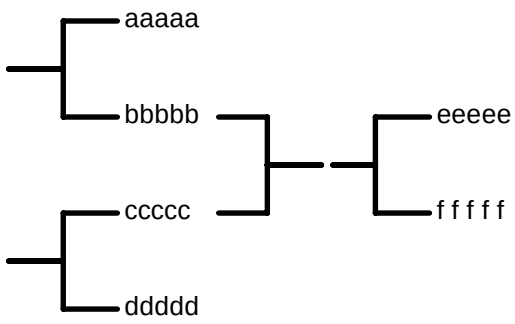


Figure 5 Disjunctive system

Several conventions have been followed. A lexeme is indicated by block capitals. Thus DEAN is a lexeme. An attestation is indicated by the use of double inverted commas. Thus “dean” (*Gwreans an Bys*: line 340) is an attestation. A word type is indicated by italics. Thus *dean* is a word type. A translation equivalent is indicated by single inverted commas. Thus ‘man’ is a translation equivalent of “dean” (*Gwreans an Bys*: line 340). A dictionary headword is indicated by bold type. Thus **DEAN** (PDMC) and **dēn** (NCED) are dictionary headwords. A grapheme is indicated by angled brackets. Thus <ʒ> is a grapheme.

1.3 Principal findings

Lexicographical history and tradition define the alphabet that is used, the alphabetical order of the macrostructure, the choice of grammatical form used as the base form, and the fields that constitute the lemma. Since lexicographical history and tradition play such an important part in the way in which Cornish lexicography is practised today, it is necessary that a survey of

lexicographical history and tradition is undertaken prior to lemmatisation of the corpus. The first stage in the process of lemmatisation is tokenisation at the rank of lexical item. Lexicon based tokenisation is to be preferred over character based tokenisation because it copes with the ranks of morpheme, word and multi-word lexeme. Following tokenisation, lemmatisation basically involves of two operations: the generation of the base form, and the disambiguation of homographs. Concerning the first of these operations, base forms may be generated from attested word types with the help of system networks. Concerning the second operation, the most reliable criterion for disambiguating homographs is grammatical difference. A lemma containing three fields, the canonical form, the part-of-speech and a semantic field label, is sufficient to provide a unique code for every lexeme attested in the corpus. Computer lemmatisation programs are not usually fully automatic with 100% accuracy, though they provide an extremely useful aid to lemmatisation. In theory at least, it ought to be possible to write a program that would lemmatise a corpus with 100% accuracy. However, the level of linguistic detail that would need to be incorporated in such a program would require that the corpus first be lemmatised before the program could be written. A solution is provided by programs with which humans interact during the lemmatisation process, thus allowing the lemmatisation database to be bootstrapped as lemmatisation takes place. The lemmatisation database may be at least partially created by means of computerised morphological processing; this is more effective when the corpus is available in normalised orthography. Computer programs can be used to automatically disambiguate at least some of the most common homographs. The macrostructure for both sides of a

bilingual Cornish-English and English-Cornish dictionary can be generated by means of interlingual lemmatisation. Interlingual lemmatisation also provides the means to identify translation equivalents and to find example sentences for each lemma.

2 Cornish Dictionaries, Glosses & Lexicographical Notes

It is essential to take stock of what has already been achieved in the field of Cornish lexicography, in order to ascertain what remains to be done. A variety of reference sources provide information about the Cornish lexicon over a period of approximately a thousand years. Glosses in the margins of Latin manuscripts give Cornish equivalents for items in the text. Glossaries provide lists of items with their translation equivalents. The notes and essays of philologists explore an assortment of data concerning lexical items. Published and unpublished dictionaries give more comprehensive accounts of the Cornish lexicon. Since lexical description is distinct from grammatical description, which is concerned with the more general rules governing a language, this discussion will not include grammatical reference sources. Although dictionaries and glossaries of dialect English provide a source for lexicographers working with the Cornish language, they fall into a different category from purely Cornish lexicographical sources. They are not included, therefore, in this discussion. Cornish dictionaries, glosses and lexicographical notes may essentially be considered from two angles; firstly from a historical perspective and secondly within a framework of typology.

2.1 The historical perspective

It is essential first to examine the history of Cornish lexicography in order to understand how the process of lemmatisation relates to the Cornish language. Lemmatisation may be seen from a social as well as cognitive perspective

when we consider the history of Cornish lexicography. Cornish lexicography has not only served to provide lexical explication; it has also evolved to develop social norms. The manner in which Cornish lexicography has been practised has been determined by the circumstances in which Cornish lexicography has taken place. During the eighteenth century, Cornish antiquarian scholarship provided the environment in which Cornish lexicography was on the whole undertaken. The broader backdrop of Celtic studies provided the setting for Cornish lexical investigation during the late nineteenth century. Subsequently in the twentieth century, the driving force for Cornish lexicographical activity was language revival. Translation has been the focus of Cornish lexicography throughout history; even onomastic dictionaries focus on the translation of Cornish names into English. Cornish lexicography has undergone three stages. During the first stage, including the early glosses and the *Vocabularium Cornicum* (VC), Latin is the target language and Cornish is the user's first language. The second stage, commencing with Richard Symonds (1644) vocabulary, is purely descriptive. In other words, the data collected by the lexicographer is merely noted down. The provision of translation equivalents supplies the meaning of lexical items. This overlaps with the third stage, in which the lexicographer partially reconstructs the lexicon. In the eighteenth century gaps in the Cornish lexicon are first filled by Lhuyd (AB) and by Borlase (VCBL) who borrow from Welsh. They are followed and in the twentieth century, by Morton Nance (ECD2, ECD3, NCED, CED) who includes many words adapted from Welsh and Breton.

Our earliest Cornish lexicographical sources date back to around the end of the 9th century A.D.. The earliest known source is a Cornish glossary in *Smaragdus's Commentary on Donatus*. This is a treatise on Donatus written by Smaragdus, abbot of Mihiel, in Latin. It contains nineteen glosses which were originally thought to be Breton. Loth (1907a, 1907c) identified them as Cornish.

Three more Cornish glosses are to be found, written on a Latin text of the "Book of Tobit", in *Oxoniensis Posterior* which dates from the 10th century. Zeuss (1853: 1060-3) mistook them for Old Welsh Glosses; Stokes (1879: 21) correctly identified them as Cornish. Figure 6 shows the Cornish gloss "depena" ('behead').

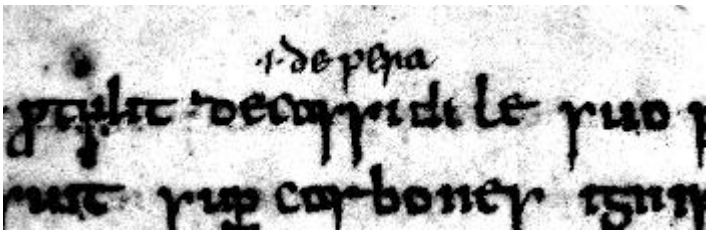


Figure 6 Gloss from *Oxoniensis Posterior*

There are a small number of Cornish glosses and phrases in the *Prophetia Merlini* by Joannis Cornubiensis. The only known manuscript of the *Prophetia Merlini* is the one in the Vatican (Vatican Cod. Ottobonianus Lat. 1474) which is a copy (Stokes 1876-1878: 85-86). It is thought that the original by Joannis Cornubiensis was written between 1153 and 1154 (Curley 1982: 222-223). Fleuriot (1974) concludes that the Cornish language fragments in the *Prophetia Merlini* are of a date that precedes the differentiation of Cornish and

Breton.

The *Vocabularium Cornicum* (VC), also known as the *Cottonian Vocabulary* and the *Old Cornish Vocabulary* is thought to date from around 1100 A.D.. It is apparently based on the earlier "English-Latin Lexicon" (ELL, St. John's College, Oxford, 154 MS. O; *cf.* AAOELG) of Aelfric, Abbot of Eynsham (c. 955 - c. 1010). Aelfric's glossary consists of a list of Latin words with their English translation equivalents appended, presumably intended as an aid to learning Latin vocabulary. Aelfric's glossary contains mostly singular nouns in the nominative case and a few adjectives. Approximately one hundred years later Aelfric's English was replaced by Cornish thus creating a Latin-Cornish glossary. It has been suggested that the Cornish is a translation of the English (Fudge 1982: 7). However this does not necessarily follow. The creator of this Cornish version, whilst recognising the value of Aelfric's original, may have been more familiar with Cornish and Latin than with English. The point is an important one; since absolute equivalence between languages cannot be taken for granted, we need to know whether the Cornish is closer to the Latin or the English. The arrangement is thematic and begins with God. The first entries are:

Deus omnipotens	'duy chefuidoc'	'almighty God'
Celum	'nef'	'heaven'
Angelus	'ail'	'angel'
Archangelus	'archail'	'archangel'

The vocabulary then continues through the stages of the creation; star, sun, moon, world, earth, sea and mankind. Then follow the parts of the body, the

ranks of the church, members of the family, crafts and their implements, animals and plants, and household goods. The *Vocabularium Cornicum* (VC) contains a total of 961 entries, compared with 1,269 in the St. John's College manuscript of Aelfric. Occasionally two translation equivalents for the Latin are given. These are linked by "uel" meaning 'or'. Examples include "broder uel braud" ('brother'), "cos uel caus" ('cheese') and "douer uel dur" ('water'). The first of each pair is Cornish and the second Welsh. The vocabulary is preceded by a calendar containing many Cornish words and the lives of Cornish and Welsh saints. The manuscript was for some time classified as Welsh since it appeared by the Latin title *Vocabularium Wallicum*. According to Lhuyd (AB: 222), the manuscript was brought to his attention by a certain John Anstis who felt that the classification as Old Welsh was inaccurate. Lhuyd confirmed that it was in fact Cornish. There is a copy with a few comments, dated 1753, of the *Vocabularium Cornicum* (VC) made by the Rev. Dr. Jeremiah Milles, Dean at Exeter Cathedral) amongst the Borlase manuscripts in the Royal Institution of Cornwall (*Mems. Of the Cornish Tongue*). Zeuss (1853: 1065-81) includes an edition of the vocabulary in his *Grammatica Celtica*. This includes a useful commentary consisting of mainly Welsh and Breton cognates and notes in Latin. Norris (1859a: Vol. II 311-435) appended an edited version to his *Ancient Cornish Drama*. This is a Cornish-Latin-English alphabetically arranged version, in which the Cornish lemma is followed by its number in the *Vocabularium Cornicum* (VC), then the page number in the *Vocabularium Cornicum*, the Latin word, its English translation equivalent, Zeuss's note (Zeuss 1853: 1065-81), and finally Norris's own remarks. This was subsequently made use of by Morton Nance (NCED).

Graves (1962) published the vocabulary with Breton and Welsh cognates appended.

Richard Symonds (1644), a lawyer serving in the Royalist army, wrote a *Diary of the Marches of the Royal Army during the Great Civil War* in which he lists twenty-four nouns in Cornish and English, the numerals from one to twenty-one and four short phrases. Long (1856: 74) observes that the preceding page of the manuscript appears to have been torn out. It is possible, therefore, that there was originally more material on Cornish. Symonds records his Cornish in the section of his diary that covers the period of his stay in Cornwall. He may have received the Cornish that he recorded from a Cornish speaker serving in the Royalist army.

In 1660, there appeared a book entitled,

> A Battledoor for Teachers and Professors to learn Singular and Plural: wherein is shewed forth by grammar or scripture examples how several nations have made a distinction between singular and plural... and in this is set forth examples of the singular and plural about thou and you in several language, Englishe, Latin, Italian, Greek, Hebrew, Caldee Saxon, Welch, Mence, Cornish, French and Spanish by George Fox, John Stubbs and Benjamin Furley.

Eighteen examples of Cornish singular and plural are included. Fox, one of the authors, was founder of the Society of Friends and had spent some time on missionary work in Cornwall between 1655 and 1666. He travelled at least as far as St. Ives and Marazion and it is possible that his travels in Cornwall provide the source for the Cornish in this book.

An anonymous Cornish-English vocabulary in the National Library of Wales (Bodewryd MS 5) is thought to date from around 1700 AD (Hawke 2001: 86).

The manuscript consists of two single-sided paper folios. There are 60 entries in all. The first page confines itself to parts of the body. The second page consists of a mixture of words and phrases in no particular order. There are a small number of words not found in any other extant sources.

The Celtic philologist, Edward Lhuyd (b.1660 – d.1709) was possibly the first qualified scholar to make a serious study of the Cornish language. In fact he spent four months in Cornwall, in 1700, learning Cornish. His informants were mainly John Keigwin, the Rev. Henry Ustick, James Jenkins and Nicholas Boson. Lhuyd had originally intended to include a Cornish-English vocabulary in his *Archaeologia Britannica* (AB). However, since the book turned out to be longer than he had expected, he postponed the publication of his Cornish vocabulary, *Geirlyfr Kyrnweig* (GK), until the second volume. Nevertheless Volume I of *Archaeologia Britannica* (AB: 41 *ff.*) contains "A Comparative Etymology" and "A Comparative Vocabulary of the Original Languages of Britain and Ireland". The "Comparative Etymology" includes "Parallel Observations relating to the Origin of Dialects, the Affinity of the British with other Languages, and their Correspondence to one another." In the "Comparative Etymology" Lhuyd (AB: 3) notes the semantic differences between cognates of the various Celtic languages. For example he observes that *Tâd gwyn* in Welsh means a step father, but in Cornish *Taz gwydn* means *a grandfather*. The "Comparative Vocabulary" is a Latin-Welsh-English-Breton-Irish-Cornish-Scots Gaelic vocabulary. The first entry is: "A, ab, abs; W. o, ygan, iurth; *From, by.* Arm. Digant; Ir.ó, a, úa. C. a, Uorth." Entries are arranged alphabetically by their Latin lemma. Participles, adverbs derived

from adjectives, as well as other derivatives and compounds are omitted.

An important feature of Lhuyd's work is his orthography. He devised his own phonetic script, based on an extended Latin alphabet. Lhuyd calls this script "The General Alphabet" (AB: 2). Lhuyd writes, "Where letters are wanting, nothing seems more natural than to borrow them out of that ancient language that is of the nearest affinity". Diacritics are also used. Gendall (1991: ix *ff.*) gives a detailed account of Cornish pronunciation based on Lhuyd's system.

Unfortunately the second volume of *Archaeologia Britannica* (AB), containing his *Geirlyfr Kyrnweig*, never appeared, due to Lhuyd's tragic death at the Ashmolean Museum in 1709. After his death, Lhuyd's manuscripts disappeared. Several years later, however, his *Geirlyfr Kyrnweig* was discovered in the National Library of Wales (*cf.* Davies 1939; Morton Nance n.d.). This consists of a small notebook consisting of 172 pages of which 162 form the vocabulary. The entries are written in black and red in Lhuyd's own handwriting, with many alterations and crossings out. The *Geirlyfr Kyrnweig* does not employ the General Alphabet that Lhuyd devised for *Archaeologia Britannica* (AB); though he does make occasional use of diacritics (circumflex to indicate a long vowel and oblique accent to indicate irregular stress). In the *Geirlyfr Kyrnweig*, Lhuyd also uses a special long-tailed-U character (see Figure 7) corresponding to < ẏ > found in *Archaeologia Britannica* (AB).

Many of the entries in the *Geirlyfr Kyrnweig* begin with three dots, <∵ >.

According to Lhuyd (AB), he obtained most of his knowledge of Cornish from manuscripts of the dramas, provided by Sir Jonathan Trelawny, Bishop of

Exeter. The *Vocabularium Cornicum* (VC), identified by Lhuyd as Cornish, provided him with another source; and words taken from the *Vocabularium Cornicum* are marked with a dagger symbol. A third source were his field notes made during his stay in Cornwall. Lhuyd (AB) admits that he made use of his native Welsh whilst attempting to recover what he could of Cornish. Morton Nance (1923) criticises Lhuyd's methodology, saying "Had that other great man, Edward Lhuyd, trusted to unlearned but habitual Cornish speakers more than to amateur philologers like John Keigwin, his four months in Cornwall might have been spent to even better purpose."

Figure 7 Lhuyd's long-tailed-U

At around the same time as Lhuyd was working on Cornish, William Hals (born 1655 – died 1737), of Fenton Gymps, was compiling *An Lhadymer ay Kernou - The Interpreter of Cornwall* (LK). Tonkin (1738) criticises Hals as being not very fluent in Cornish and suggests that before publication *An Lhadymer* should be "carefully revised by some learned discreet persons". Tonkin also regretted informing Lhuyd of Hals' vocabulary since he felt it had been instrumental in preventing Lhuyd from publishing his own vocabulary. Morton Nance (n.d.) describes the work as "an attempt by one who knew next

to nothing of Cornish to impose on others who knew even less". Hals' vocabulary was never published, a part (LK), running from A to BLIGH, may be found in the National Library of Wales. There is also a copy in the *Gwavas Manuscripts* (59r to 78v) which runs from A to CLUID.

William Gwavas (1676 - 1741), of Gwavas in the parish of Sithney near Helston, Cornwall, was a barrister and compiler of a collection of Cornish songs, verses, proverbs and letters. The *Gwavas Manuscripts* (119v-125r) include his Cornish-English glossary. Gwavas's glossary contains 271 entries and is arranged alphabetically under the first letter of the head word and runs from ABEM to OZE YOUNK. George Borlase (1733) of Penzance made a copy of some of Gwavas's papers. He gives the following account of pages 150b-163b.

> The following Book conteyning A great many Cornish words and their Etymologicall Significacions was written by Mr Wm. Gwavas of Newlyn in Mountsbay and delivered me to be transcribed in the yeare 1733.
>
> (Borlase 1733)

Although George Borlase's copy contains very little that is not found in other Gwavas and Tonkin manuscripts, it is not an exact transcript of any known Gwavas manuscript.

Thomas Tonkin (born 1678 – died 1742) was born at Trevaunance, St. Agnes, Cornwall. He obtained a degree at Queen's College, Oxford, and then settled on the family estate at Lambrigan in St-Piran-in-the-Sands. Tonkin (1736) suggested to his friend William Gwavas that they publish a Cornish vocabulary. The proposed vocabulary was not published in Tonkin's lifetime,

however the manuscript (CLEV) can be found amongst the Cornish manuscripts in Bilbao, Spain (*Bilbao Manuscripts*). It contains words in both Middle and Modern Cornish forms. Lhuyd's General Alphabet is used for many of the entries and Lhuyd may have been the source for these. Gendall (SDMC: iii) suggests that, Lhuyd's *Archaeologia Britannica* (AB) cannot have been the source, since the item **iutîziou**: 'justices', which is misspelt in the *Archaeologia*, is correctly spelt, **iustîziou**, in Tonkin's vocabulary (CLEV). Of course, this does not necessarily follow since Tonkin may simply have been correcting an error that he discovered in the *Archaeologia*.

Dr. William Borlase published the second edition of his *Observations on the Antiquities Historical and Monumental, of the County of Cornwall* in 1769. This included what he described as "a Vocabulary of the Cornu-British Language" (VCBL), which amounts to fifty pages containing approximately 4,000 entries in total (Borlase, William 1769: 413-64). Borlase does not use the Greek characters that Lhuyd used to extend his alphabet. He does, however use some diacritics. Among the Borlase manuscripts (*Mems. of the Cornish Tongue*; *cf.* Jenner 1912) are copies of manuscripts by Lhuyd, Gwavas, Tonkin, Ustick, Scawen and Boson, in Borlase's handwriting. He also obtained a copy of the *Vocabularium Cornicum* (VC) from the Rev. Dr. Jeremiah Milles. These represent his sources and are all acknowledged as such in the preface to his vocabulary, where mention is also found of a "Baxter's Glossary", a "Davies's Dict." and the "J.T.Tregere MS.". The first of these is the *Glossarium antiquitatum britannicarum: sive syllabu etymogicus antiquitatum veteris Britanniae atque Ibe temporibus Romanorum / auctore*

Willielmo Baxter ... accedunt ... Eduardi Luidii ... . De fluviorum, mont urbium, &c. in Britannia nominibus, adversaria posth (GAB2) that was published in 1733. The second of these is John Davies' *Antiquae linguae britannicae, nunc vulgò dictae Cambro-britannicae, a suis cymraecae vel cambricae, ab aliis wallicae et lingvae latinae dictionarium duplex. Prius, britannico-latinum, plurimis venerandæ antiquitatis Britannicæ monumentis respersum. Posterius latino-britannicum. Accesserunt Adagia britannica, & plura & emendatiora, quàm antehàc edita* (ALB) that was published in 1632. It is not clear whether the last is the same Tregear that translated Edmund Bonner's *Homelies* into Cornish. The few words for which Borlase acknowledges J.T. Tregere as the source, are not found in Tregear's *Homelies*.

Borlase examined Lhuyd's papers in the library of Sir Thomas Seabright, Bart.. Although Lhuyd's *Geirlyfr Kyrnweig* (GK) was not among them, Borlase did find an "imperfect" English-Cornish vocabulary, "...and in other scattered memorandums, I found several Cornish words I had not seen before, which in this work are inserted..." (Borlase 1769: 413). Borlase filled in gaps in the Cornish vocabulary by borrowing Breton words given by Lhuyd in his *Archaeologia Britannica* (AB). This marks the first attempt at reconstructive Cornish lexicography. Borlase's manuscripts also include "First Essay for an English-Cornish Vocabulary" (*Mems. of the Cornish Tongue*: Part I, 26-43). Lhuyd (AB: Tit. VIII) forms the basis of this, the principle being for Welsh and Breton translation equivalents to be appended to the Cornish lemma. The Welsh items are mostly present, but the Cornish and Breton are unfinished. "Cornish Words digested under two Initials with their English: Ludgvan, 8

April, 1749" (*Mems. of the Cornish Tongue*: Part II, 1-92) is a rough of copy that forms the basis of his Vocabulary of the Cornu-British Language (Borlase 1769). It consists of two alphabetical series of entries, the second being a supplement. With regard to the "Vocabulary of the Cornu-British Language", Gendall (SDMC: iii) is of the opinion that:

> The fact of its containing borrowings, inventions, misprints and misunderstandings does not detract from the value of much of its contents which, again by comparison with the work of other writers, as also from our knowledge of vocabulary survivals into the 19th and 20th centuries, can be seen to be correct.

In 1790, William Pryce published his *Archaeologia Cornu-Britannica* (ACB). Prince Louis Lucien Bonaparte (1861), when he uncovered Tonkin's and Gwavas' original letters in the 1860s, accused Pryce of plagiarism, asserting that Pryce took the unpublished vocabulary and notes of Tonkin and Gwavas together with Lhuyd's grammar and published the entire collection together without acknowledgement. Bonaparte alleges that the original manuscript, now in the provincial library at Bilbao, Spain (*Bilbao Manuscripts*), shows the work, published by Dr. William Pryce (ACB) as his own and without acknowledgement, to have been compiled in 1730 by Tonkin or Gwavas. The story concerning this alleged plagiarism has been perpetuated by Jenner (1925: 425) and Berresford Ellis (1974: 136). However, in the preface of the *Archaeologia Cornu-Britannica* (ACB), Pryce clearly acknowledges his use of the manuscripts of Tonkin and Gwavas.

> In this collection Mr Tonkin took the lead, being determined to publish a Cornish Word-Book in his then proposed History and Antiquities of Cornwall illustrated, in three volumes quarto.... he died before he had compleated the work. He left, indeed, a large mass of MS. books, but they were thrown together without any sort of order or connection.... Mr. Tonkin was assisted in his undertaking by the critical knowledge and industry of William Gwavas, Esq. who was indefatigable in collecting and ascertaining words for his use and arrangement.... In consequence of the death of Mr. Tonkin, this collection... was taken into the protection of the late Robert Hoblyn, of Nanswhidden, Esq. .... It was afterwards taken thence, and committed to my trust by favour of the late John Quick, Esq. ... who, with reiterated expressions of his wish to see it warmed into life, consigned it to my care for correction, additions, and publication; to which end I pledged my diligence and application, with whatever assistance I could procure from the MSS. before mentioned, together with some detached papers from Mrs. Veal, the daughter of Mr. Gwavas; from Mrs. Mary Ustick, the widow of the Rev. Henry Ustick, of Breage; and from the papers of Mr. John Bosons, of Newlyn. I also applied to Miss Foss, the representative of her grandfather Thomas Tonkin, Esq. for the use of his other MSS. to which I had access, and from which I extracted all that I could find valuable in that rich mass of indigested materials....
>
> (ACB: n.p.)

Pryce made an important contribution to Cornish lexicography by publishing his *Archaeologia Cornu-Britannica* (ACB). The vocabulary contains approximately 4,000 entries. Like Lhuyd, Pryce uses a dagger symbol to indicate items found in the *Vocabularium Cornicum* (VC). He gives separate entries to homographs and head words are often found in their inflected and/or mutated forms. Williams (LCB) describes Pryce's work as full of errors and he goes on to say that he "felt satisfied that Pryce was entirely ignorant of the Cornish language". However, as Bonaparte (1866) points out, since the work in question was compiled by Gwavas or Tonkin, the criticism would have to apply to them, and they could scarcely be said to be 'entirely ignorant of the Cornish Language'. Pryce was certainly aware of Johnson (DEL) when he compiled his *Archaeologia Cornu-Britannica* (ACB) and cites Johnson's well

known definition of a ‘lexicographer’ as a ‘harmless drudge’ in his preface.

In 1808, the Rev. Richard Polwhele published his *Cornish-English Vocabulary* (CEV) containing approximately 2,200 entries. Like Borlase, he uses the standard Roman alphabet, with certain diacritics. He acknowledges his sources as Borlase, Pryce and a “large collection of words from Whitaker’s invaluable papers”. This is possibly John Whitaker (born 1735 – died 1808), Rector of Ruan Lanihorne, who wrote an unpublished manuscript history of the parish of Ruan Lanihorne.

Charles Rogers of Stonehouse, Plymouth, compiled a “Vocabulary of the Cornish Language” in 1861 (VCL, Bodleian MS Cornish d 1). Rogers’ vocabulary has never been published. His sources include Norris’ (1859a) transcription of the *Ordinalia*, Davies Gilbert’s transcriptions of *Pascon agan Arluth* (Gilbert 1826) and *Gwreans an Bys* (Gilbert 1827), Borlase (*Mems of the Cornish Tongue*; VCBL), Pryce (ACB), Polwhele (CEV), Whitaker (1804), Tonkin (CLEV) and Carew (1602).

In 1865, the Rev. Robert Williams of Rhydycroesau, Wales, published his *Lexicon Cornu-Britannicum - Gerlyvr Cernewec* (LCB). At the time, this represented the most thorough and comprehensive Cornish dictionary to date, containing approximately 8,000 entries, covering all periods of Cornish. Williams was a Welsh speaker. The dictionary contains many citations from the texts - with line references, English translation equivalents, and Welsh, Breton, Irish, Gaelic and Manx cognates. Williams follows the unusual procedure of giving a separate insertion to each variant spelling of a word. He

also attempts to solve the problem of variable orthography by amalgamation. These reforms, which include diacritics, the adoption of Lhuyd's <DH> for voiced <TH>, and the substitution of <C> for the letter <K> in all cases, have met with a mixed response. Dr. Whitley Stokes (CG2: 138) criticises Williams dictionary, saying that "Mr. Williams has throughout his *Lexicon* been misled by Welsh analogy." Williams' dictionary was similarly criticised by Prince Louis Lucien Bonaparte (1866) and Professor Joseph Loth (1902b: 236) and more recently Richard Gendall (SDMC: iii). Furthermore Stokes (CG2: 138) is critical of Williams' orthography, writing that analogy with Welsh misled Williams into distinguishing between <DH> and <TH>.

In 1868 Whitley Stokes published a "Cornish Glossary" (CG1); this was intended to provide a supplement to Williams' *Lexicon* (LCB) and contains approximately 2,000 words, most of which are not included in Williams' *Lexicon* and some of which represent corrections. Stokes (CG2: 137) emphasises that the known sources have not been exhausted in the search for lexis. Stokes' sources include the *Domesday Book* (Cornwall), *Pascon agan Arluth* (Stokes 1861), the *Ordinalia* (Norris 1859a), *Gwreans an Bys* (Stokes 1863), *Archaeologia Britannica* (AB).

In 1887 Frederic Jago published his *English-Cornish Dictionary* (ECD1). His sources include Stokes (CG1) and *Beunans Meriasek* which Williams was not able to include in his *Lexicon* (LCB). His aims were as follows:

> 1. To collect all the words which should find a place in an English - Cornish Dictionary.
>
> 2. To quote some Cornish phrases for the sake of illustration.

> 3. To give the various forms or spellings of the words just as they are found in the remains of ancient Cornish, without constructing a single word, or phrase, and without alteration or addition.
>
> 4. To place the various spellings of the words in a gradational form, for the sake of their being more easily compared.
>
> 5. To give one authority at least for each word and phrase, for the sake of an easy reference to the originals.
>
> (ECD1: xiii-xiv)

Jago (ALDC: x) considered Cornish dialect English to be an important source:

> even now the Cornish people are speaking a large number of Celtic or ancient Cornish words without being aware of it. The Cornish dialect may well be the shadow, or penumbra, of the ancient Cornish language, the link between the old and the new tongue, between Celtic and English.

He, therefore, included many words derived from dialect English, though in his opinion, "Some of these are doubtful, but it is safer to keep them than to lose them" (ECD1: xiv). This is reminiscent of Schuchardt's (1866-1868: III: 35) 'substratum' theory in which he posits that gradual modification of a *lingua franca* towards a pidgin results from continual interaction between the 'substratum' languages of the Etruscans, Iberians and Celts, and the language of their conquerors. Jago (ECD1) gives a separate entry for each inflected form of the verbal paradigm. He is also thorough in giving all the variant spellings of an item, its attestation and examples of usage.

The manuscript of the play *Beunans Meriasek* was discovered at Peniarth Library in 1869 (National Library of Wales, Peniarth 105), and, in 1900, Stokes (GCDBM) published a glossary to the play, containing 2000 previously unattested Cornish words, in the *Transactions of the Philological Society*.

The end of the 19th century saw a growth in interest in Celtic Studies with several journals, that occasionally include articles relating to the Cornish language, commencing publication. *Revue celtique* the first of these began publishing in 1870. It was followed in 1897 by the first volume of *Zeitschrift für celtische Philologie* and the first volume of *Archiv für celtische Lexikographie* in 1900. These journals are an invaluable resource to the Cornish lexicologist. However, scattered as they are amongst several journals, these articles on Cornish can be quite difficult to track down. I shall, therefore, list the principal articles here.

Between 1870 and 1932, the journal, *Revue celtique*, published several articles relating to Cornish lexicology, including "The Manumissions of the Bodmin Gospels" (Stokes 1870-1872), "Cornica" (Stokes 1876-1878; Stokes 1879-1880), "Les gloses de l'Oxoniensis posterior sont-elles corniques" (Loth 1893a), "Les mots 'druic', 'nader', dans le Vocabulaire cornique" (Loth 1893b), "Etudes corniques" (Loth 1897, 1902a, 1902b, 1903, 1905), "Remarques et corrections au *Lexicon Cornu-Britannicum* de Williams" (Loth 1902b), "Le cornique: à propos d'un livre de M. Henry Jenner" (Loth 1906), "Cornoviana" (Loth 1911a, 1911b, 1913), "Questions de grammaire" (Loth 1914, 1917-1919) and "Contributions à l'étude des textes corniques" (Cuillandre 1931, 1932).

The *Zeitschrift für celtische Philologie* published a small number of articles relating to Cornish between 1897 and 1982. These articles include "A Welsh (Cornish?) Gloss in a Leyden MS" (Lindsay 1897), "The Preverbal Particle 're' in Cornish" (Williams 1910), "Is Cornish Actually Dead" (Allin-

Collins 1930), “Cornish Words in the Tregear MS” (Morton Nance 1954), “Celtic Manuscripts in Spain and Portugal” (Hull 1958-1959) and “Notes corniques” (Quentel 1982).

Between 1900 and 1907 the *Archiv für celtische Lexikographie* published five articles relating to Cornish. These include Stokes’ (GCDBM) “A Glossary to the Cornish Drama ‘Beunans Meriasek’ ”, Stokes’ (1900b) corrections to Norris’ (1859a) *Ancient Cornish Drama,* Loth’s (1900) transcriptions of Boorde’s (1555) “Colloquies” and *William Bodinar’s Letter*, Loth’s (1907a) article identifying the glosses on Smaragdus’ Commentary on Donatus (Paris Bibliotheque Nat. MS.Lat. 13029) as Cornish, and Loth’s (1907b) article concerning various etymologies.

Five articles relating to the Cornish language appear in *Études celtiques* between 1938 and 1991. These articles include “Review of R, Morton Nance’s (NCED) *A New Cornish-English Dictionary*” (Vendryes 1938), “Middle Welsh, Cornish and Breton Personal Pronominal Forms” (Hamp 1958-1959), “Les fragments du texte brittonique de la *Prophetia Merlini*“ (Fleuriot 1974) and “The Nouns Suffixes *–ter*/*-der*, -(*y*)*ans* and *–neth* in Cornish” (George 1991).

By the 1920s, interest in Cornish as a revived language was steadily growing. However learners were experiencing difficulty not only with finding new words to express modern concepts, but with the many discrepancies of spelling. Robert Morton Nance (1929) devised a standardised spelling system which became known as ‘Unified Spelling’. According to Berresford Ellis

(1974: 155), “Morton Nance learnt his first Cornish from Borlase’s (1769) *Antiquities* and Sandys’s (1846) *Specimens of Cornish Provincial Dialect*”. Morton Nance’s dictionaries that followed were based on his new spelling and are not so much descriptive as reconstructive. Prior to Morton Nance, lemma lists had included variant spellings and mutated forms. In Morton Nance’s dictionaries the canonical forms that constitute the lemma list are first properly established.

By the 1930s the *Federation of Old Cornwall* Societies had grown so much that it was able to sponsor a new dictionary. This was to establish fixed spellings and paradigms of verbs. The preparation for the press was done by Arthur Saxon Dennett Smith and, in 1934, Morton Nance and Smith published *An English-Cornish Dictionary* (ECD2). Morton Nance and Smith introduced words borrowed from Breton and Welsh and respelled them according to what they considered their most likely Cornish form. These borrowings are marked in the dictionary with an asterisk. T. Eurwedd Williams added a Welsh section to Morton Nance and Smith’s ECD2 to create a trilingual *English-Cornish-Welsh Dictionary* (ECWD) in two volumes. However this has unfortunately remained unpublished. The manuscript resides in the National Library of Wales (MSS.12514 and 12515).

Robert Morton Nance’s *A New Cornish-English Dictionary* (NCED) was published in 1938 by the *Federation of Old Cornwall* Societies. The £2,000 that paid for the publication of this dictionary was raised by public donation. Morton Nance described this as his “life work”. The aims (NCED: Introduction) were to include every known word of Cornish and to

include many words presumed to have formed part of the language and to provide an acceptable standard spelling. Morton Nance's Unified Cornish is based on the Middle Cornish of the *Ordinalia* and *Pascon agan Arluth*. Morton Nance (NCED: Introduction) regarded these texts as representing "the most perfect form of the language as well as the best known". George (GKK: 6) observes that the NCED attempts two tasks: "to act as a glossary for all words found in traditional Cornish literature, and to provide a lexicon for revived Cornish".

The existing texts provided the main source for the NCED. However, place-names, as spelt in medieval documents especially, and dialect English supplied many more. In addition, gaps in the lexicon were filled in by respelling Welsh and Breton cognates to allow for phonological differences. Occasionally borrowings were taken from Middle English. Borrowings are marked in the dictionary with an asterisk (*). Middle-Cornish words, however re-spelt, have no distinguishing mark. Those respelt from Old-Cornish (older than 1300) are marked with a dagger symbol (†) and those respelt from Late-Cornish (later than 1600) are marked with a double-dagger symbol (‡). Reconstructions of the many missing genders, plural forms, infinitive-endings and verb paradigms were made by Morton Nance by analogy with Breton and Welsh. In this matter Breton was felt to be closer to Cornish.

Apart from English translation equivalents, Morton Nance's NCED includes sources, examples of usage and idioms for many of the words. Paradigms of verbs and pronominal prepositions are confined to appendices. Actual spellings and variants are added in brackets, although Lhuyd's (AB)

General Alphabet is represented in ordinary type. Quotations are given in Unified Cornish either to illustrate idiomatic usages or to amend old translations. Variant and contracted Middle-Cornish forms are given, with reference to which Morton Nance (NCED: iii) states, “the form first given being usually preferable, even when it differs from that most usual”. Word combinations that are translated by one word in English are hyphenated. Text references are restricted to less common words. Until the 1990s the NCED remained the most modern work on Cornish in existence. Morton Nance’s own heavily annotated working copy can be found amongst the documents in the Morton Nance Bequest in the Courtney Library of the Royal Institute of Cornwall in Truro.

John Tregear’s Cornish translations of the homilies from Bonner’s *Profitable and Necessary Doctrine* (Tregear n.d.; Bonner 1555) were discovered in April 1949 by John Mackechnie amongst some papers of the Puleston family of Wales, in the British Museum. The following year (1950), Morton Nance published “Cornish Words Occurring in Tregear MS” (CWOT). This glossary contains 50 entries, which not only included fresh words but confirmed or corrected some conjectural genders, plurals and infinitive endings.

In 1952 Morton Nance published his *English-Cornish Dictionary* (ECD3). Richard Gendall prepared the first draft for Morton Nance’s editing. Morton Nance and Smith’s (ECD2) *English-Cornish Dictionary* formed a basis, but in addition Richard Gendall put into reverse order Morton Nance’s NCED of 1938. All previous dictionaries had relied on the printed additions of the texts. The ECD3 profited by Morton Nance’s consultation of photostats of

the original manuscripts. An additional source were Tregear's Cornish translations of Bonner's homilies, which were unknown when NCED was published in 1938. In ECD3, the judicious development of neologisms replaces some of the borrowings from Breton and Welsh.

In 1955 Morton Nance published another *Cornish-English Dictionary* (CED). This included a few adaptations and neologisms from the ECD3 of 1952 and omits the vast majority of comparative and historical material to be found in the NCED of 1938.

Berresford Ellis (1974: 194) points out that modern Celticists such as Jackson largely ignore Morton Nance's dictionaries and quote their Cornish from Williams' (LCB) *Lexicon Cornu-Britannicum* of 1865. However, Morton Nance had access to the researches of Joseph Loth and Whitley Stokes. More accurate transcriptions of the texts than Williams used were available in Morton Nance's time. And sources hitherto unavailable for study in Williams' time, including *Beunans Meriasek*, the *Charter Endorsement* and various manuscripts by Lhuyd, Borlase, Tonkin and Gwavas, were used by Morton Nance. As a result, he achieved a far greater degree of accuracy than did Williams.

Professor Charles Thomas (1972), of the Institute of Cornish Studies, criticises the basis of Morton Nance's Unified Spelling:

> Our Institute takes the view that the so called Unified Spelling invented by Morton Nance has never been explained, i.e. we have never had any real discussion of the principles on which it was based. We regard the dictionaries with their high proportion of words invented by the comparative method as suspect, because they don't give dated forms, and we feel that some of the lost words can probably be recovered from dated Middle Cornish place-names and may prove to be other than the forms invented for them by Morton Nance. Lastly, following the work of the Leeds Survey of English Dialects, we suspect that the pronunciation currently used for modern Cornish (based on an ultimate form of Wessex Middle English) may be wrong and that the true phonetic range is still just recoverable from an area west of an isogloss that cuts off the Land's End and part of the south side of the Lizard.

In 1980, Andrew Hawke began work on a historical dictionary of Cornish. The basis for this consists of a lexicographical index, a bibliographic index, a manuscript archive and a text archive. The lexicographical index provides access to all the most important published and unpublished dictionaries and lexicographical notes on Cornish. Dictionaries were photocopied on different coloured paper, for identification. Individual entries were then cut out and affixed to A6 sheets of paper to form an alphabetical card index. Morton Nance's Unified spelling was used for the lemma list. The bibliographic index was to include any publication that refers to Cornish or to a particular Cornish word. By noting all the Cornish words referred to, a lexical index as well as a bibliography would be compiled. The manuscript archive includes microfiche copies of texts. The text archive includes texts prepared in machine readable form. Hawke planned to use *Oxford Concordance Program* software to produce concordances, linking all orthographical forms. Homographs would then be distinguished. A system of cross references would then enable every form to be found and a suitable canonical form selected. Unfortunately, this vast undertaking has not been completed (Hawke 1982).

The growing popularity of revived Cornish created a need for new words relating to aspects of everyday life in the twentieth century. Since these words neither existed in the historic vocabulary nor in the limited range of neologisms to be found in Morton Nance's dictionaries (ECD2, ECD3, NCED, CED), Snell and Morris compiled three *Cornish Dictionary Supplements* in order to meet this demand. The first of these, *Kitchen Things - On the Roads* (CDS1), appeared in 1981. The second of the supplements, *Home and Office* (CDS2) was published in 1984. The third supplement, *General Words* (CDS3), compiled by Morris alone was published in 1995.

In 1991 Richard Gendall published *A Students' Dictionary of Modern Cornish - Part 1, English – Cornish* (SDMC). This dictionary covers the Modern Cornish period, and contains approximately 9,000 English head words. Morton Nance's (1929) Unified spelling is abandoned in this dictionary, which, according to Gendall (SDMC: i), "contains every word, in every found variety of spelling, that could be gleaned from all available sources from the 16th century onwards, and all the words from the rich characteristic dialect of West Cornwall that might have a bearing upon a study of its Cornish language". Gendall acknowledges his sources for each Cornish word form, but only gives the line number for those items taken from the play, *Gwreans an Bys*. His earliest sources include, Andrew Boorde (1555), Tregear (n.d.) and *Gwreans an Bys.* His most recent sources include items taken from English dialect. Gendall (SDMC: iii) asserts that many English dialect words found in West Penwith "are descended directly from the Cornish vernacular, sometimes in a form little if at all different from that in which they may have occurred in

the living language, but at other times much altered". To illustrate this, he cites dialect words which do not appear within the corpus of written Cornish, yet have cognates in Breton and Welsh. George (GKK: 6) asserts that Gendall does "not adequately distinguish between words from the traditional Cornish language and words in the dialect of English in use in Cornwall". This accusation is unjust since Gendall clearly marks words in his dictionary that are taken from dialect, "T", which he explains:

> traditional: being material transmitted orally from 18th, 19th & 20th cent. without any part. authorship though collected by identifiable persons. Names of individuals are given where known, but informants are very numerous. T covers dialect glossaries among which are those printed in the Old Cornwall magazines, 'Cornish Provincial Dialect' by Wm Sandys, 1846, 'Glossary of words in use in Cornwall', by M.A. Courtney & T. Couch, 1880, 'Glossary of Provincial Words', by F. Jago, 1880, 'A Glossary of Cornish Words' by Joseph Thomas, 1895, 'Old Newlyn Speech', by Ben Batten, 1984, MSS collection held by the Institute of Cornish Studies.
>
> (SDMC: vi)

In 1984, Ken George completed a thesis for the degree of Doctorat du Troisième Cycle on the *Phonological History of Cornish* at the University of Western Britanny. This was followed by the publication of *The Pronunciation and Spelling of Revived Cornish* (George 1986), in which he recommends that the Middle Cornish period of around 1500 A.D. should serve as a phonological basis for Revived Cornish, and that the spelling system be adapted to provide a phonemic representation of this (George 1986: 4). In 1987, a decision was made by the Cornish Language Board to convert the Unified orthography of Morton Nance (1929) to the new orthography called 'Kernewek Kemmyn'.

In 1993, George published his *Gerlyver Kernewek Kemmyn: an Gerlyver Meur, Kernewek – Sowsnek* (GKK), with the aid of a grant from the Human Resources, Education, Training and Youth Task Force of the Commission of the European Community. The dictionary contains approximately 9,000 entries and incorporates most of the words from the first two dictionary supplements of Snell and Morris (CDS1, CDS2). George (GKK: 7) explains that "The dictionary is aimed at the speakers and learners of Revived Cornish..."; in other words, it is not primarily intended for the interpretation of the corpus of old texts. Sources include the dictionaries of Morton Nance (ECD2, ECD3, NCED, CED), Graves' (1962) thesis on the *Vocabularium Cornicum* (VC), Snell and Morris' (CDS1, CDS2) supplements, Haywood's (1982) dissertation on Old Cornish, Padel's *Cornish Place Name Elements* (CPNE), and the monthly Cornish language magazine *An Gannas*. George's GKK has been much criticised, particularly with regard to his Kernewek Kemmyn orthography (Penglaze 1994; Williams 1995, 1996, 2001; Mills 1999). In particular, Mills (1999) and Williams (2001) have shown there to be a great many inaccuracies in George's GKK.

George's *The New Standard Cornish Dictionary: An Gerlyver Kres: Cornish-English English-Cornish* (NSCD) was published in 1998. This is an abridged version of his GKK of 1993 with the addition of an English-Cornish section.

In his *A Practical Dictionary of Modern Cornish: Part One Cornish-English* (PDMC), published in 1997, Gendall standardises Cornish orthography by selecting a preferred spelling for each head word from among the forms attested in the corpus of Modern Cornish. Where Gendall has included

words attested only in Lhuyd's (AB) General Alphabet, he has respelled these in the general style of Late Cornish orthography. He similarly respells items attested only in English dialect dictionaries. Gendall (PDMC) includes many neologisms borrowed directly from English though these may also undergo respelling by Gendall.

1998 saw the publication of Gendall's *A New Practical Dictionary of Modern Cornish: Part Two English-Cornish* (NPDMC). This is essentially a reversal of Gendall's PDMC published the previous year but with further standardisation of the Cornish orthography. Gendall explains his approach to standardisation thus,

> It soon became apparent that strict adherence to found forms was giving Modern Cornish the appearance of being complicated while at the same time it was being claimed as having the advantage of simplicity. The problem has been throughout the whole of historical Cornish literature its many authors have used their own orthographies, frequently varying these during the course of one work, even of one line of writing. In reconstructing a usable idiom, differing parts of any one verb, for instance, are found scattered among the writing of more than one author, and in various orthographies, to the extent that practically nowhere is it possible to obtain all the requisite parts either in the works of a single writer or in a single spelling system.
>
> The only sensible solution to this problem has been to standardize, within reasonable limits, and mainly as regards roots and terminations, while at the same time ensuring not only that any standardizations are in line with the most typical examples to be found, but that as far as practicable words remain otherwise unaltered. In practice, there is no facet of standardization that will not be found in an example of at least one historical author, so that standardization means nothing more than a reasoned choice from what is available.
>
> Gendall (NPDMC: A)

Nicholas Williams' (2000) *English-Cornish Dictionary: Gerlyver Sawsnek-Kernowek* (ECD4) gives Cornish translation equivalents in Williams' (1997)

Unified Cornish Revised orthography. This orthography is Williams' revision of Morton Nance's (1929) Unified Cornish orthography and according to Williams (1997: 5) is based on a corpus of three texts close to each other in date of composition: Tregear's (n.d.) *Homilies*, *Beunans Meriasek* of 1504, and *Gwreans an Bys* of 1611. With more than 24,000 head words, William's ECD4 contains more entries than any previously published Cornish dictionary. This is the result of the inclusion of a vast number of 20th century neologisms. Unfortunately Williams does not give any sources for his Cornish translation equivalents. It is, therefore, not possible to determine from this dictionary which of these neologisms have been taken from well attested 20th century usage and which have simply been created by Williams for the purpose of enlarging his dictionary. Williams' inclusion of a vast number of neologisms has been criticised by Kennedy (2001:316):

> … do we not compromise the particularity of Cornish by devising a neologism to translate every word in English? The bilingual dictionary has a bearing on this, its dual-columns serving as a sort of DNA template for the lexical reconstruction of Cornish in the image of English. Where the English entry has no corresponding Cornish equivalent we are tempted to devise one. Thus Williams has: creativity: creaster, invisibility: anweladewder, linear: lynek, libertarian: lybertarek, internationalize: keskenedhlegy. Such one word solutions to perceived lexical gaps subtly change the character of Cornish. It certainly alters the thought-world of Cornish to one in which such concepts as creativity suddenly exist. Of course we need answers but perhaps these should include partial solutions, not the tidy insistence on neat semantic equivalents. This requires recognition that Cornish brings the benefit of different perspectives and subjectivities and that this has implications for vocabulary. … the cumulative effect of his neologisms is … a radical reshaping of Cornish. … it does feel as though some of the particularity of the language is lost with every act of modernization and expansion.

2.2 Onomastic dictionaries

The steady stream of onomastic dictionaries produced from the start of the eighteenth century onwards is the result of a fascination with Cornish place names and personal names. The main concern of these dictionaries is to supply the etymologies of place names and personal names. Assonance forms the basis of the etymologies supplied by these Cornish onomastic dictionaries. The etymologies are thus for the most part conjectural. In the twentieth century, the capricious spelling of attestations came to be seen as an obstacle to the systematic analysis Cornish of onomastic terms. One solution was to give onomastic terms only in normalised spelling. Alternatively the attested form might be glossed with its equivalent form in normalised spelling. Usually information regarding the pronunciation of onomastic terms is not given. In view of the fact that, from their written form, the pronunciation of Cornish place names is often obscure, this appears to be a serious omission.

The rather dubious inferences given in dictionaries of place and personal names have been frequently denounced. William Borlase (1749) remarks that "etymology gives great latitude for imagination and conjecture". Reaney (1960 :1) maintains that "more nonsense has been written on place names than on any other subject except, perhaps, that of surnames". Morton Nance (1963a: 1) writes,

> A book to contain every recorded place-name of Cornwall, with all the changes, some quite beyond imagination, that they have undergone, and especially with a meaning past dispute given to each one, is a work which we will never see.

Morton Nance, furthermore, doubts that a dictionary of Cornish place names

could be produced in which one could have confidence in more than half the definitions. Dictionaries of place or personal names fall into two categories; those listing names and those that list elements of names.

Eighteenth century onomastic glossaries are relatively small in terms of their number of entries. In his notebook of 1710, William Gwavas (*Common-Place Book of William Gwavas*) included some place name etymologies. Gwavas also wrote a Cornish - English vocabulary entitled “An Essay towards an Alphabeticall Etymologicall Cornish Vocabulary with ye signification thereof in English of the names of persons places Towns fields Tinworks & rivers &c.” The manuscript contains approximately eighty entries and mentions two existing manuscripts, one by Thomas Tonkin and the other by William Hals. It is unpublished and a copy made by Borlase (*Mems. of the Cornish Tongue*: Part II, 127-8) can be found in the Cornwall Records Office (*cf.* Jenner 1912: 167). Amongst the manuscripts of William Borlase (*Mems. of the Cornish Tongue*: Part II, 128-53) is “An alphabetical List of the Principal Places in Cornwall with their signification in the Cornish Tongue”. Pryce’s (ACB) *Archaeologia Cornu-Britannica* (ACB) of 1790 includes “An Alphabetical List of the Cornish British Names of Hundreds, Parishes and Villages in Cornwall, according to The Ancient and Modern Orthography, and expressive of their Locality and contingent Circumstances”. This is comprised of glossaries of 32 place name elements, the nine hundreds, 63 parishes and approximately 800 villages. The definitions that Pryce attempts should be treated with great caution.

Nineteenth century onomastic dictionaries are more comprehensive

than their eighteenth century counterparts. In 1870, R.S. Charnock published his *Patronymica Cornu-Britannica* (PCB) which gives a fairly exhaustive list of approximately 1,500 Cornish surnames. Entries are listed alphabetically and include comments about the etymology and meaning of the name. Dexter (CN: 69), however, cautions that Charnock's derivations should be treated with caution. According to Charnock (PCB: vi), his sources include lists from a Miss Hext (sister of J.H. Hext, late of Gray's Inn) and a Mr. J.C. Hotten, publisher. However most of his material was gleaned from the Post Office Directory for Cornwall, Pryce (ACB) and Polwhele (1816). In 1871, J. Bannister published his *Glossary of Cornish Names* (GCN). This contains approximately 20,000 place names. Dexter (CN: 69) cautions that "Many of the derivations must be most carefully considered before adoption".

In 1926 Dexter published his *Cornish Names* (CN) in which he attempts to explain over 1,600 Cornish names. Dexter purports that the book is not merely a glossary but "an attempt to show the reader how he can interpret many more for himself". It is therefore not alphabetically arranged but is divided in to chapters dealing with certain themes, such as; natural features, works of man, and foreign influences. There is an alphabetical index at the back of the book. In 1928 J. Gover completed his *Place Names of Cornwall* (PNC), which contains approximately 10,000 entries with attempted explanations of their etymology. Unfortunately this has remained unpublished. In 1945, Robert Morton Nance published two articles in *Old Cornwall* entitled "Celtic Personal Names of Cornwall", in which he discusses etymologies and possible meanings.

In 1954, R.R. Blewett of St. Day began the task of analysing the 1953 Electoral Registers for the five Cornish Parliamentary Divisions. From a list of approximately a quarter of a million voters, he determined the number of entries bearing each surname and their distribution in Cornwall. Perceiving that many were associated with place names, he compared his list with Gover's *Place Names of Cornwall* (PNC). After consulting dictionaries to unravel meanings, in 1970, he completed *Celtic Surnames in Cornwall, their Distribution and Population in 1953, their Origins, History and Etymology* (CSCDP) in two volumes. This has remained unpublished.

Circa 1960, Robert Morton Nance published *A Guide to Cornish Place Names* (GCPN) containing a list of approximately 650 place name elements. These are listed in alphabetical order, in Morton Nance's (1929) Unified spelling. The problem for the user is that the forms in which they occur on the map in place names can vary quite a lot from Morton Nance's Unified spelling.

In 1970 Bice published his *Names for the Cornish - 300 Cornish Christian Names* (NC). His aim was "to help Cornish parents with the practical business of choosing distinctively Cornish Christian names for their children" and is essentially revivalist in nature. His sources include the 1327 Lay Subsidy Rolls, the miracle plays, official documents and parish registers, place names, and the *Bodmin Gospels*.

In 1972, G. Pawley White published *A Handbook of Cornish Surnames* (HCS1), which contains the names of Celtic origin with derivations, areas of concentration in 1953 and relevant place names. Blewett's CSCDP formed the

basis of this. Pawley White (HCS2: 4) mentions “a manuscript collection of Cornish names with their possible derivations by the late Edwin Chirgwin...” as one of the sources for the second edition of his handbook. Other sources used by Pawley White include the Lay Subsidy Rolls of 1327 and 1523, and the Parish Registers from 1600-1812.

In 1973, P.A.S. Pool published *The Place Names of West Penwith* (PNWP1). His sources include Gover (PNC), Charles Henderson, Robert Morton Nance, Parish Tithe Apportionments and various other manuscripts. The second edition (PNWP2) published in 1985 also draws on the Penwith Hundred Court Rolls and includes an additional section on names of natural features such as hills and headlands. Included in the lemma list are the names of all farms and other settlements in West Penwith. Pool lists the names alphabetically as they are found on current Ordnance Survey maps.

In 1983, Julyan Holmes published *1,000 Cornish Place-Names Explained* (TCPNE). The glossary of place names is alphabetically arranged. The lemma consisting of the place name is as it is spelt today on the Ordnance Survey map. This is followed by a transcription into Morton Nance’s (1929) Unified spelling and then an English translation. Holmes does not explain how he arrived at his interpretations.

In 1985 Oliver Padel published his book, *Cornish Place-Name Elements* (CPNE). This is not a dictionary of place names, but a dictionary of the elements that constitute Cornish place names. The dictionary contains approximately 800 such elements. The collections of the Cornish Place-Name

Survey, at the Institute of Cornish Studies, form its basis. These collections are largely founded on those of Gover (1928), Charles Henderson and the Rev. W. Picking. Padel draws from parallel place name material in Welsh and Breton, in particular for the period before c.1200 A.D., which he takes to be a period when the three languages formed almost a linguistic unity. The forms for the lemma list are mostly drawn from attested Middle Cornish forms. This poses a problem, since firstly not all place name forms are attested in the Middle Cornish texts and secondly spellings are not consistent in the Middle Cornish texts. Padel deals with the first problem by either creating a hypothetical Middle Cornish form, which he marks with a star, or else he borrows an Old or Modern Cornish form. The second problem he attempts to solve by arranging words according to "their intended sounds, rather than literal spelling" (CPNE: xvii). In this manner <C> and <K> are treated as a single letter, but vocalic <Y> is treated separately from consonantal <Y>. Together with the fact that many Cornish place name elements are not found today in their Middle Cornish form this makes the dictionary rather difficult to use.

In 1988 Padel published *A Popular Dictionary of Cornish Place Names* (PDCPN) which gives an account of just over 1,000 Cornish place names. The names which form his lemma list are taken from the 1982 *Ordnance Survey Quarter-Inch Map* of Cornwall. These are sequenced alphabetically. The name is followed by the map grid reference, the name of the parish in which the place is located, the earliest recorded form of the name and other spellings, a suggested meaning, alternative names by which the place has been known, and occasionally an incorrect derivation is refuted.

In 1990, Pool published *The Field-Names of West Penwith* (FNWP). The lemma list is not exhaustive, consisting of a selection of field names found in West Penwith, "chosen to illustrate the use of the Cornish Language in field names and to cast light on the history and topography of West Penwith" (FNWP: 28). Crofts, moors, commons and areas of waste ground are included as well as fields proper. Pool's sources date from the 17th century. He also makes use of the Tithe Apportionments of circa 1840 for comparison with earlier records. Lemmata are arranged in alphabetical order in the left hand column. In the right hand column the name is transcribed in Morton Nance's Unified spelling and an English translation is suggested. Explanations which Pool regards as 'doubtful' are qualified as 'probably' or 'possibly'.

In Weatherhill's *Cornish Place Names and Language* (CPNL), published in 1995, we find the entries grouped together according to their region. In all, this book includes nearly 2000 Cornish place names. Thus there is a chapter on the "Place Names of Penzance, St. Ives & Lands End", another on the "Place Names of Helston & Lizard", and so on. Within each chapter the place names are then listed alphabetically. For many of the entries, Weatherhill includes a pronunciation field and frequently he gives one or more etyma.

Several dictionary compilers have discovered that assembling a lemma list of name elements is confounded by the deviations in spelling with which they occur. Attempts to amalgamate these, however, make the list difficult to use, since it no longer includes all the forms as they naturally occur. If a place name dictionary is to have popular appeal, there is pressure on the compiler to give etymologies and definitions. However such explanations should be

treated with the greatest caution.

2.3 Interlingual relations

Fundamental to the process of reconstruction is the notion of borrowing from Welsh and Breton. The practice of appending Welsh cognates to Cornish items goes back to the *Vocabularium Cornicum* (VC) in which a small number of Welsh cognates have been included. Sebeok (1962: 365), in his typology of dictionaries of the Cheremis language, describes the relationship between the components of each entry as follows:

> Within an entry, the object language may be represented by a (5) single form or by multiple forms. If the object language is represented by multiple forms, the relationship between them may be of two kinds: (6) based on form - a dictionary of cognates - ...; or (7) based on meaning - a dictionary of synonyms...

Lhuyd recognises Sebeok's distinction and provides both a "Comparative Etymology" (AB: 3), which brings together cognates of the various Celtic languages and notes semantic differences between them, and a "Comparative Vocabulary" (AB: 41 *ff.*) in which entries are arranged alphabetically by their Latin lemma. This has the effect of bringing items with the same meaning together. That Lhuyd (AB), in his "Comparative Etymology", notes the semantic differences between cognates of the various Celtic languages is important, since it is not always appreciated by modern speakers of Cornish that such differences exist. Consequently they all too often attempt to interpret Cornish by analogy with Welsh and Breton. The English-Cornish dictionary type brings together both the cognates and the synonyms of Cornish. Both Morton Nance (ECD2) and Gendall (SDMC) made use of this feature to begin

their lexicographical endeavours.

Adopting a modern spelling system for the canonical forms in the lemma list, such as Morton Nance's Unified Cornish or George's Kernewek Kemmyn, has its problems for the user. In the case of place name dictionaries, the spelling of place name elements, as found on street signs or the ordnance survey map, can vary quite a lot from these modern spelling systems. A similar situation exists with regard to the miracle plays and other texts which make up the corpus of traditional Cornish. Although these represent the sources for the dictionaries of Morton Nance and George, the forms to be found in these texts are not the same as the canonical forms given in the dictionaries.

All the dictionaries considered so far are either bilingual or multilingual; in other words they are translation dictionaries. According to Kromann, Riiber and Rosbach (1991: 2713), the typology of translation dictionaries may be considered from three viewpoints.

> The user aspect is concerned with the target group of dictionary users, their needs and competence, and the kinds of situation that occur.
>
> The empirical aspect includes the establishment of relevant text corpora and the excerpting of lexical units.
>
> The linguistic aspect is concerned with equivalence relations between fields of lexical units in the language pair and the paradigmatic and syntagmatic relations of these fields.

With regard to the user aspect, we do not know who were the writers and readers of the glosses on *Smaragdus's Commentary on Donatus* and *Oxoniensis Posterior*, and the *Vocabularium Cornicum* (VC). Nevertheless, if these early lexicographic endeavours were intended to help the learner of

Latin, whose first language is Cornish, then it would seem likely that the writer had first language intuition of Cornish. In other words, Latin is the object language. All subsequent sources, from Lhuyd (AB) onwards have Cornish as the object language. In other words, they are intended for those whose first language is English. Furthermore, these later compilers did certainly not have first language intuition of Cornish. In the twentieth century the dictionaries of Morton Nance (ECD2, ECD3, NCED, CED) and George (GKK) are intended for speakers and learners of Revived Cornish.

With regard to the empirical aspect there are principally four sources; firstly previous dictionaries and glosses, secondly a corpus comprised of miracle plays and other texts, thirdly dialect and fourthly place names and personal names. All Cornish dictionaries depend on earlier dictionaries, borrowing from their inventories of words and taking over the translation equivalents that they provide. Lhuyd (AB) made use of the *Vocabularium Cornicum* (VC). Borlase (VCBL) made use of the vocabularies of Lhuyd (AB), Tonkin (CLEV) and the *Gwavas Manuscripts*. Lhuyd's vocabulary (AB) is the source for much of Tonkin's vocabulary (CLEV). The vocabularies of Borlase (VCBL) and Tonkin (CLEV) provided the sources for Pryce's (ACB) vocabulary. This process continued into the twentieth century with Morton Nance (ECD2, ECD3, NCED, CED) making use of previous lexicographical sources. Most recently George's (GKK) sources include the dictionaries of Morton Nance (ECD2, ECD3, NCED, CED) as well as Snell and Morris' (CDS1, CDS2, CDS3) supplements.

The Cornish miracle plays, various poems and songs and other scraps

have also provided a source. Sir Jonathan Trelawny, Bishop of Exeter provided Lhuyd with access to manuscripts of the dramas. Williams (LCB) cites from *Pascon agan Arluth*, the *Ordinalia* and *Gwreans an Bys*. To these sources Morton Nance (NCED) adds *Beunans Meriasek*, the Modern Cornish *Gwavas Manuscripts*, Lhuyd's *Archaeologia Britannica* (AB) and Pryce's *Archaeologia Cornu-Britannica* (ACB). When working from such sources the lexicographer is faced with the problem of deciding precisely what constitutes a lemma. The writers of the miracle plays were not consistent in the marking of word boundaries by a space. And there are certain clitics and elisions to be found among the Cornish texts, which add to the confusion. The lemma lists of earlier lexicographers have inevitably influenced those who have followed.

Transcribers of the medieval texts do not agree on the precise interpretation of the orthography of the originals. Lexicographers since Lhuyd (AB) have made attempts to standardise their own orthographies. Lhuyd (AB) devised his own phonetic spelling which he called "The General Alphabet". Williams (LCB) made efforts to standardise spelling by the amalgamation of forms. Morton Nance (1929) developed the spelling system known as Unified Cornish and most recently George (GKK) has introduced a phonemic spelling system known as Kernewek Kemmyn. Gendall (SDMC) sidesteps the issue of providing a Cornish lemma list by compiling an English - Cornish dictionary. This brings together both cognates and synonyms under a single English lemma.

In 1887 Jago (ECD1) was the first to use Cornish dialect English as a source. This has been a source for Cornish dictionaries ever since and in

particular for Gendall's (SDMC) *Students' Dictionary of Modern Cornish*. Another source has been place names, notably in the dictionaries of Morton Nance and George. George (GKK), in fact, includes nearly one hundred extra words gleaned from Padel's *Cornish Place Name Elements* (CPNE).

There have been a number of attempts to reconstruct the lexicon where there are gaps. Lhuyd (AB) borrowed from his native Welsh, and since the lexicographers that followed borrowed from Lhuyd this needs to be borne in mind when their works are appraised. Borlase (VCBL) borrowed from Breton to fill gaps in the vocabulary. The twentieth century saw the revival of Cornish and with that, a need was created for many new words to express modern concepts. Morton Nance (NCED) included many words presumed to have formed part of the language. In other words, if cognates exist in both Welsh and Breton it was assumed that Cornish must also have possessed a cognate. Gaps in the lexicon could, therefore, be filled by respelling Welsh and Breton cognates to allow for phonological differences. Landau (1989: 78) calls words that could exist but for which no record exists to prove that they have ever been used 'latent words'. Morton Nance also made occasional borrowings from Middle English. In Morton Nance's ECD3 of 1952, neologisms started to creep in. The dictionary supplements of Snell & Morris (CDS1, CDS2, CDS3) introduced a large number of neologisms many of which were adopted by George (GKK).

With regard to the linguistic aspect, it is the bilingual or multilingual nature of all glossaries and dictionaries of Cornish that have been produced so far, that is the chief concern. Palmer (1981: 87) notes that hyponymy relations

vary from language to language. The number of lexical units in a given lexical field, therefore, also vary. As a result, whilst the meaning of a lexical unit in one language may be related to the meaning of a lexical unit in another language, the equivalence relation of the subsenses in the two languages can vary.

George (GKK: 17) complains about Morton Nance's giving "a large number of meanings, even to words which appear only once in the texts...", and he adopts the policy of supplying not more than three translation equivalents per entry. In this matter George fails to distinguish between what he calls a "meaning" and a translation equivalent. According to Catford (1967: 130), a translation equivalent may be defined as "a target-language text, or item in text which changes when and only when a given source-language text or item is changed". With regard to sense, it may be true to say that if an item occurs only once in the corpus that, unless the citation in question is ambiguous, that particular item in that particular context carries only one sense. However, even if a source-language item is attested only once in the corpus, and, therefore, presumably has only one sense, there may be a number of target-language translation equivalents that convey that sense. In the following example, the item "ben" could be translated by either summit or top.

> "A lene yn hombronkyas vghell war ben vn meneth" (*Pascon Agan Arluth*: stanza 16)
>
> 'Thence he led Him high on the summit / top of a mountain',

This does not indicate that in this particular context "ben" has two senses. But it does illustrate that the one sense expressed by "ben" in this context may be

expressed by either ‘summit’ or ‘top’ in English. Catford (1967: 133) notes that, although translation equivalence may be established between sentences, it may be more difficult to do so between individual items. Nida (1958: 281) asserts that,

> (1) No word (or semantic unit) ever has exactly the same meaning in two different utterances; (2) there are no complete synonyms within a language; (3) there are no exact correspondences between related words in different languages.

Consider the Cornish verb **resek**, George gives one English translation equivalent, ‘run’. *The Oxford Advanced Learner’s Dictionary of Current English* (OALD4: 1107-8) gives 32 senses for the verb ‘run’ (not including phrasal verbs). Are we to assume, therefore, that the Cornish **resek** can also be used to convey all of these senses? Surely not! It would be safer to say that **resek** is a translation for ‘run’ in certain circumstances. In other words **resek** and ‘run’ do not cover the same semantic range.

Lexicographical sources suggest a number of translation equivalents for the Cornish word **pen**. In Figure 8, the horizontal columns represent the lexicographical source, and the vertical lines the translation equivalent given in the lexicographical source.

	VCBL	*ACB*	*LCB*	*ECD1*	*CED*	*Brown (1984)*	*GKK*
beginning			+	+	+		
chapter					+		
Chief			+	+	+	+	
conclusion			+				
End		+	+	+	+		+
extremity			+				
Head	+	+	+	+	+	+	+
promontory	+						
summit			+	+			+
Top				+	+		
upper part			+				

Figure 8 Equivalents of Cornish PEN

The English translation equivalents ‘chief’, ‘conclusion’, ‘end’, ‘extremity’, ‘head’, ‘summit’, ‘top’ and ‘upper-part’ and in addition one more, ‘supreme’ (making a total of 9) are confirmed by the following citations taken from the Corpus of Cornish (Mills 1992: ch. 8).

chief:

> “del osa dev thy’n ha **pen”** (*Passio Domini*: line 732)
>
> ‘As You are God to us and chief’,

end, conclusion:

> “ef a sef the **pen** try deth” (*Resurrexio Domini*: line 52)
>
> ‘He shall rise again at the **end / conclusion** of three days’.

head:

> “war ow **fen** curyn a spern lym ha glev” (*Resurrexio Domini*: line 2581)
>
> ‘upon My **head** a crown of thorns sharp and piercing’

supreme:

> “gylwys o why . **pen** arlythy” (*Resurrexio Domini*: line 325)
>
> ‘You are called **supreme** lords’.

summit, top:

> “A lene yn hombronkyas vghell war **ben** vn meneth” (*Pascon Agan Arluth*: stanza 16)
>
> ‘Thence he led Him high on the **summit / top** of a mountain’

In addition to these, Mills (1992: ch. 8) also identifies three idiomatic usages of **pen**.

Kettep pen: *every one.*

> “me a genes yn lowen ha’m dyscyblyon kettep **pen** the’th arhadow” (*Passio Domini*: line 460)
>
> ‘I will go with thee gladly, and my disciples **every one**, at thy bidding’

Pen pusorn: *refrain/chorus of a song.*

> “ha ty tulfryk **pen pusorn** dalleth thy’nny ny cane” (*Resurrexio Domini*: line 2353)
>
> ‘and do thou, Tulfryk, begin to sing for us a **refrain/chorus** of a song.’

War pen deulyn: *kneeling.*

> “pup-oll war **ben y dheulyn**, a gan yn gordhyans dhodho” (*Pascon Agan Arluth*: stanza 245)
>
> ‘everyone, **kneeling**, will sing in worship to Him.’

It may be concluded that there is no justification for limiting the number of translation equivalents given to three as George (GKK) does, other than that George happens to feel that three should be the maximum, and such restriction is not rational, warranted or helpful.

There are essentially three possible equivalence relations; full, partial and

zero. Morton Nance (NCED: 20) gives one English translation equivalent, ‘hand’, for the Cornish lexical item **lüf**. If hand may be used to translate **lüf** in all its possible contexts, then it is a full translation equivalent. In other words it covers the whole range of the lexical meaning of the lemma. However it cannot be presumed that the equivalence relation is the same in both directions. Morton Nance (ECD3: 80) gives two Cornish translation equivalents, **lüf** and **dorn**, for the English lemma ‘hand’. **Lüf** and **dorn** are therefore partial translation equivalents of hand. In this event there is a divergence from English to Cornish and convergence from Cornish to English. So the equivalence relation depends on the direction of translation. (For a full analysis of the Cornish items **lüf** and **dorn** see Mills (1992: ch. 6). There are parallels here with other Celtic languages. For example, in Gaelic LÀMH donates the ‘hand and the arm’ whilst DORN donates a ‘fist’.

Kromann, Riiber and Rosbach (1991: 2716-7) argue that an equivalence relation may be posited between a lemma and its translation equivalents or between the individual meanings of the lemmatised word and the particular meaning of the translation equivalent word. This distinction leads to very different results. For example, Morton Nance (NCED: 174) gives ‘to owe’, ‘deserve’, ‘pay’, ‘be worth’, ‘avail’, ‘requite’, ‘repay’, and ‘recompense’ as translation equivalents for the lexical item **tylly**. Partial equivalence results between units if the equivalence relation is established between the lemma and its translation equivalents. On the other hand, if subsenses of the lemma **tylly** were postulated and the equivalence relation were to be established between these and corresponding subsenses of the translation equivalents, then full

equivalence would result.

Zero or surrogate translation equivalents provide approximate translation in cases where the object language lexeme is culture specific. For example, George (GKK: 189) gives the English translation equivalent ‘violin’ for the Cornish lexeme, **krowd**. The word **krowd** is attested in the miracle play, *Origo Mundi* (line 1997) of the late 14th century. Musical instruments have changed a great deal since the 14th century and a Cornish **krowd** from this period is not the same instrument as a twentieth century violin. So ‘violin’ is a surrogate translation equivalent for the Cornish **krowd**. Kromann, Riiber and Rosbach (1991: 2718) identify the need for a precise encyclopaedic explanation to accompany such a surrogate translation equivalent, something which George (GKK), in fact, fails to give.

Kromann, Riiber and Rosbach (1991: 2720) note that comments such as field markers, encyclopaedic explanations, etc. not only provide useful information in themselves, they also serve as aids to meaning discrimination. For the active dictionary user, encoding in Cornish, meaning discrimination is essential. Examples may assist with meaning discrimination. Gendall (SDMC: 8) discriminates between the senses of the English lexeme BANK by incorporating field markers in brackets (see Figure 9). However, he gives no indication to help the dictionary user to choose between the various forms given within each field, and no examples are given.

bank (rampart etc.) bankan,tụban(L),(river) gladn,glands,gland(T), terneụan,terneụan an aụan,torneụan an aụan(L), ladn(B), (commerce) tryssor(B).

Figure 9 SDMC, English lexeme BANK

Palmer (1981: 3-4) observes that dictionaries

> provide definitions by suggesting words or phrases which, we are given to understand, have the 'same' meaning, though what is sameness is a problem that we shall not be able to escape. The extent to which meaning is dealt with in terms of equivalence of terms is even more clearly brought out when we deal with foreign languages.

It is clear that the bilingual approach to Cornish lexicography that we have seen so far, has not provided an thorough semantic analysis of the Cornish lexicon. Only where the translation equivalents given can be taken to be absolute, can we be sure that meaning is unambiguously conveyed. The many occurrences of partial or surrogate equivalence present problems of disambiguation.

In order to establish relations of equivalence between the individual meanings attached to the lexical units of a pair of languages, Baldinger (1971) proposes a full semantic analysis of both languages. This has never been done for the Cornish language. In fact, the semantics of Cornish has only ever been described in terms of translation equivalents. Kromann, Riiber and Rosbach (1991: 2714) propose that "a bilingual lexicography with any claim to scientific rigour must establish and maintain its own representative corpora in accordance with the nature of the target groups the projects are aiming at".

It is necessary to stipulate what a semantic analysis of Cornish entails.

According to Leech (1981: 208), “The SEMANTIC SPECIFICATION (or definition) of a word is a representation of its meaning in terms of componential or predicational analysis ....” Palmer (1981: 110) also notes that componential analysis “allows us to provide definitions for all these words in terms of a few components.” Leech (1981: 206) further suggests that special formal language is the only entirely adequate means of representing the meaning of a lexical item. However he notes that such formal language would convey little to the average dictionary user and so the lexicographer has to resort to paraphrase. In other words the lexicographer does not give the sense of the lemma. Instead s/he provides another expression which shares the same sense as the lemma.

It is this writer’s opinion that a semantic analysis of the Cornish lexicon should determine semantic fields, determine semantic relations between items, and identify syntagmatic relations such as collocations, idioms, and multi-word lexemes. There are no speakers of Cornish today with first language intuition of Cornish which could be used to extract semantic analyses from computer generated concordances. Instead, explicit criteria which derive from the corpus itself are needed for the semantic investigation of the lexicon.

Mills (1992: ch. 10) identifies a number of explicit criteria that the semanticist may invoke when working from computer generated concordances. Positive-negative frames can be elicited from a corpus by producing a concordance of negative particles. These frames determine that the meanings of a pair of items do in fact contrast, and secondly they serve as a means of highlighting the significant differences in two events and the relationship of entailment

between them. Causal frames may be elicited by producing a concordance of clause connectors that signal cause. These can reveal relations of entailment. Part of speech may distinguish homographs and derived senses. In the case of Cornish, the rules governing mutation of initial letters is especially useful in determining part of speech. Case roles and valency impart very explicit information about the lexis they involve. For example, in the sentence Peter ate the fish, two roles can be identified; the EATER and the EATEN. A concordance of things EATEN provides a list of co-hyponyms which may be defined as `types of food' or given the semantic marker EDIBLE. Significant collocates may be generated automatically by some concordancers, such as TACT, these frequently reveal a semantic field or more generally a topic shared with the keyword. Collocation within a given semantic-field may indicate the contiguity of a set of items. Alternatively contrasts between items of a set may be indicated by the collocations that they form individually with discrete semantic fields. Indirect illocutions or speech acts may distinguish various senses of the item under investigation. Anaphora may provide further information about an item under investigation. Paratactic listing reveals items that share semantic fields. If these are found in pairs they are often antonyms. Paratactic lists may be elicited from a corpus by producing a concordance of conjunctions.

The above list is by no means exhaustive. The idea of producing concordances of certain signals, such as negative particles, clause-connectors or conjunctions, could be extended. A concordance of intensifiers, for example, would reveal items that could be labelled ATTRIBUTES. Procedures of this

kind could extract a considerable amount of information about a large number of lexical items relatively quickly and easily. Furthermore this information would be explicitly supported by attestation in the corpus.

3 The Corpus of Cornish

The set of all dictionary sources, which are themselves of several types are referred to as the 'corpus'. Some writers (Hausmann and Wiegand 1991: 337; *cf.* Pan Zaiping and Wiegand 1987: 234 *ff.*) prefer the term 'dictionary basis'. Nowadays corpora are stored electronically. There are many different types of corpus. A general corpus consists of a body of texts which provide the basis for a general dictionary. A general corpus needs to be balanced so as to contain texts from a variety of genres as well as samples of both spoken and written language use. A monitor corpus is one which is kept up-to-date by the continuous addition of new material and the deletion of old material. Specialised corpora deal with a specific genre or text type, such as child language, dialect, or scientific text. The types of corpora so far mentioned are sometimes referred to as synchronic corpora since they represent language at a particular time. As such, they contrast with diachronic corpora which represent language over a long period.

The Corpus of Cornish is comprised of texts from the Middle Cornish and Modern Cornish periods. The corpus thus covers a period ranging from the late 14th century to the latter part of the 18th century with, in addition, a couple of tiny fragments from the 19th century. An understanding of the nature and characteristics of these texts is essential. The diachronic range encompassed by the corpus is the reason for a great deal of the variation in orthographic practice. As well as this diachronic variation, even within a single document considerable evidence of capricious spelling is found. Published critical

editions of the source texts vary in their reliability. It was thus judged necessary to make new critical editions for the electronic corpus. Above all, a new lexicon based tokenisation is incorporated in these new critical editions.

The corpus is comprised of texts from a relatively small number of informants, particularly during the Middle Cornish period. The Middle Cornish component of the corpus is comprised of six texts, only one of which bears a colophon to identify its author. The extent to which each individual Middle Cornish text is the work of a single author is uncertain. Little is known about the authors of the Middle Cornish texts; it is even possible that they were not mother-tongue speakers of Cornish. By contrast, a far greater number of informants are represented in the Modern Cornish component of the corpus. For quite a few of the Modern Cornish informants biographical particulars are known, including whether or not they spoke Cornish as a mother-tongue. It might be felt that some informants are more reliable than others, such as those known to be mother tongue speakers of Cornish. Such knowledge might cause a lexicographer to have a preference for a particular attested base form to serve as the canonical form.

3.1 Chronology of the corpus of Cornish

The Corpus of Cornish upon which this project is based is diachronic. Historical Cornish is traditionally divided into three phases Old Cornish, Middle Cornish and Modern Cornish (Berresford Ellis 1974; George 1986: 8-10; SDMC: ii). Because so few historical Cornish texts are known to exist, all extant material has been included in the corpus. This results in a corpus that is

not balanced quantitatively concerning diatextual features and concerning diachronic representation. A somewhat restricted range of genres are to be found in the Middle Cornish component of the corpus. Practically the entire corpus of Middle Cornish is in verse and concerns Christian religious topics. Two miracle plays form most of the Middle Cornish component of the corpus which is thus written to be spoken. A much wider variety of genres is to be found in the Modern Cornish component of the corpus. The corpus of Modern Cornish thus includes examples of reported conversation, plays, prayer and liturgy, sermons/homilies, proverbs and sayings, prophecies, stories, poetry, verse, prose, epigrams, short rhymes, song lyrics, elegies, epitaphs, monumental inscriptions, letters, biblical translations, and mottoes.

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* (VC), a Cornish Latin glossary. Figure 10 lists the items of the Old Cornish phase.

Cornish glosses in *Smaragdus's Commentary on Donatus*	End of 9th century
Three Cornish glosses in *Oxoniensis Posterior*	10th century
Manumissions in the *Bodmin Gospels*	10th century
Vocabularium Cornicum (VC)	circa 1100
Cornish glosses in *Prophetia Merlini* by Joannis Cornubiensis (John of Cornwall)	12th century

Figure 10 The corpus of Old Cornish

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. Figure 11 lists the items of the Middle Cornish phase.

The *Glasney Cartulary*	13th century quoted in a 15th century manuscript
The *Charter Endorsement*	late 14th century
Pascon Agan Arluth	15th century
The *Ordinalia*	circa 1500
Beunans Meriasek	1504
Black Book of Merthen	1506-1536

Figure 11 The corpus of Middle Cornish

The Modern Cornish phase lasts 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 (AB) important account of the sound system and grammar. Figure 12 lists the items of the Modern Cornish phase.

1.	*Star Chambers*	1547
2.	*Fyrst Boke of the Introduction of Knowledge* by Andrew Boorde	1555
3.	*Image of Idlenesse* by Olyver Oldwanton	1555
4.	*Tregear Homilies* translated from English by John Tregear	1560
5.	*Exeter Consistory Court Depositions*	1569-1572
6.	*The Survey of Cornwall* by Richard Carew	1602
7.	*Gwreans an Bys*	1611
8.	*The Northern Lasse* by Richard Brome	1632
9.	*Diary of Richard Symonds*	1644
10.	*Keigwin Manuscripts* 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",	
11.	Scawen "Antiquities Cornu-Britannick or Observations on an Ancient manuscript Entitled Passio Christi"	Written circa 1680 published London 1777
12.	*Gwavas Manuscripts*	
13.	The *Bilbao Manuscripts* compiled by Thomas Tonkin	
14.	*An Lhadymer ay Kernou* compiled by William Hals	
15.	*William Hals' History* Words used by Revd. John Jackman in administering the sacrament in the 17th century.	Early 18th century
16.	The *Penzance Manuscript*	
17.	Thomas Tonkin's Manuscript B	Early 18th century
18.	Thomas Tonkin's Manuscript H	Early 18th century
19.	*Archaeologia Britannica* by Edward Lhuyd (AB)	1707
20.	*Common-Place Book of William Gwavas*	1710
21.	*Jottings by William Gwavas* on reverse of a legal document	1732
22.	*William Gwavas' copy of John Boson's Pilchard - Curing Rhyme*	
23.	*William Gwavas' copy of John Boson's Ten Commandments, Lord's Prayer and Creed*	
24.	*Mems. of the Cornish Tongue* compiled by William Borlase. Copies of manuscripts by Lhuyd, Gwavas, Tonkin, Ustick, Scawen and Boson in Borlase's own hand.	1748
25.	*Manuscript of Nicholas Boson's* "Nebbaz Gerriau dro tho Carnoack", Henry Ustick's Hand	1750
26.	Scawen: "Collectanea de Cornubia", "Observations on the Tongue", sayings: "Cows Nebas Cows da ….", The Lords Prayer, The Creed (*Enys Collection*).	
27.	The *Scawen Manuscripts* (in Tonkin's Hand)	

28. William Bodinar's Letter	1776
29. *Observations on a Manuscript Entitled Passio Christi...* by William Scawen	1777
30. *Geirlyfr Kyrnweig* compiled by Edward Lhuyd (GK)	
31. Lhuyd's Phonetically Spelled Transcript of James Jenkins' Verses	
32. *Manuscript Belonging to Lhuyd* Extracts From "The Dutchess Of Cornwall's Progresse To See Ye Land's End And To Visit Ye Mount" Nicholas Boson. James Jenkins' Cornish Rhymes in Lhuyd's phonetic script	

Figure 12 The corpus of Modern Cornish

The Corpus of Cornish upon which this project is based is comprised of all the available extant texts of the Middle and Modern Cornish periods.

A central feature in the structure of the Corpus of Cornish is the dimension of idiolect. The corpus consists of written texts which may not so much represent the products of groups of speakers, but more the work of individuals. Sometimes these individuals identify themselves by means of a colophon. Sometimes the texts are anonymous. Some have sought to derive a history of phonological change from these texts (George 1983, 1984, 1986). However due to the idiolectal nature of the texts, any verdicts concerning historical change should be approached with the greatest of caution. A description is here given of all the texts that comprise the corpus of Cornish.

The *Charter Endorsement* is found *in dorso* (on the back of) a charter relating to St. Stephen-in-Brannel, dated 1340 AD. The endorsement was discovered in 1877 by Henry Jenner (1915-1916) while he was working in the British Museum cataloguing recently acquired manuscripts. Amongst these were certain Additional Charters concerning grants of land in Cornwall in the reigns of Edward III to Edward IV. These Charters had been presented to the

British Museum by Sir Charles Trevelyan in 1872. Jenner (1904: 26) dates the endorsement to around 1400 AD. However, Jenner's (1915-1916) later comment, that "the hand is not more than forty or fifty years later than that of the face of the document," suggests a date of around 1385. It is, thus, the oldest extant instance of running text in Cornish.

The manuscript of the *Charter Endorsement* contains the following 23 graphemes: <A>, <B>, <C>, <D>, <E>, <F>, <G>, <H>, <I>, <K>, <L>, <M>, <N>, <O>, <P>, <R>, <S>, <T>, <U>, <V>, <W>, <Y>, <ȝ>. It is difficult to distinguish several of the graphemes since they share the same form in the manuscript: <N>, <V> and <U>, <C> and <T>, long <S> and <F>. A long-tailed-z character, <ȝ>, represents dental fricatives, [θ] and [ð]. Legibility of the endorsement is further hindered by the poor condition of the manuscript, which is dirty and smudged. The first eleven lines are only partially readable due to blotting. These difficulties concerning the legibility of the endorsement have led to several slightly differing transcriptions (Jenner 1877, 1915-1916; Stokes 1879-1880; Morton Nance 1932, 1947; Campanile 1963; Berresford Ellis 1974: 42-43; Toorians 1991: 4-6). The text of the *Charter Endorsement* consists of 41 lines of secular Cornish verse. In the first twenty six lines, the narrator is offering in marriage a girl who is highly recommended as a good housewife. The second half of the text advises the girl how to behave with regard to her husband and how to maintain the upper hand in the marriage. The whole texts consists of 190 word tokens and 126 word types, according to my own count.

The *Glasney Cartulary* contains a single sentence of Cornish containing

the prophecy, “in Polsethow ywhylur anethow”. The word “anethow” is homographic and may be translated as either ‘wonders’ or ‘dwellings’. Thus the prophecy may be translated as either ‘in Polsethow wonders will be shown’ or ‘in Polsethow dwellings will be shown’. The *Glasney Cartulary* is thought to have been written in the 15th century (Berresford Ellis 1974: 35; Murdoch 1993: 14).

Pascon Agan Arluth, also known by its English title *The Passion Poem*, is thought, on account of its palæography, to have been written in the 15th century (ACB Preface; LCB: Preface; ECD1: viii; Murdoch 1993: 19). *Pascon Agan Arluth* is an anonymous Cornish Poem of the Passion from Palm Sunday to Easter Morning. It is taken from the four gospels with additional material from the medieval legendary *Gospel of Nicodemus*.

Thirteen manuscript copies of *Pascon Agan Arluth* are known to exist (British Library Harleian N. 1782; Lambeth Palace Library 806.2 art. 17; Cornwall County Record Office Scawen MS, Fortescue Collection; Bodleian Gough Cornwall 4; Bodleian Carte 269 art. 5; Bodleian Gough Cornwall 3; Royal Institution of Cornwall Tonkin B; *Gwavas Manuscripts*: 51-58; Bodleian Corn. c.1; British Library Add. MSS 14934; National Library of Wales Panton 74; Bodleian Corn. c.3; Royal Institution of Cornwall Gatley). The copy in the British Library (Harleian N. 1782) is thought to be the oldest of these (Jenner 1904: 26; Murdoch 1979; Hawke 1981), and Hawke (1979: 50) identifies this with the copy once owned by William Scawen. Scawen gave this copy to John Anstis who arranged for it to be translated by Martin Keigwin and his son John Keigwin. Anstis then supplied Edward Lhuyd and Bishop Jonathan

Trelawney with copies.

Gilbert (1826) produced an edition of *Pascon Agan Arluth* including Keigwin's translation based on the copy in the Archbishop's Library, Lambeth. Gilbert's (1826) edition has been much criticised for its numerous inaccuracies of transcription (LCB: Preface; ECD1: viii; NCED: iii; Hooper 1972: 3). Stokes (1861) produced an edition with a translation into English based on the oldest of the manuscripts (British Library Harleian N. 1782). There is a transcription in the Journal *Kernow* by Morton Nance (1934-1936). A transcription in Unified Cornish with English translation by Morton Nance and Smith appeared in the journal *An Lef Kernewek* This was later edited and published by Hooper (1972). Pennaod (1981) produced an edition which includes the poem in its original spelling taken from Stokes (1861) edition, Morton Nance and Smith's Unified Cornish transcription, and a translation into Breton. Edwards (1993) produced an edition that includes the poem in its original spelling based on Pennaod's (1981) edition and Morton Nance's (1934-1936) edition. Edwards (1993) also includes a transcription into Kernewek Kemmyn normalised orthography and Edwards' own English translation.

The oldest of the extant manuscripts of *Pascon Agan Arluth* (British Library Harleian N. 1782) contains the following graphemes: <A>, <B>, <C>, <D>, <E>, <F>, <G>, <H>, <I>, <J>, <K>, <L>, <M>, <N>, <O>, <P>, <Q>, <R>, <S>, <T>, <U>, <V>, <W>, <X>, <Y>, <ʒ>. The graphemes <U> and <V> are homographic as are <I> and <J>. The graphemes long <S> and <F> are also frequently homographic. The long-tailed-z grapheme, <ʒ>,

represents dental fricatives and is attested in free variation with the graphemes <DH>, <D> and <TH>.

The poem contains 259 stanzas of eight lines each. In total, the poem contains 10,091 word tokens and 2,501 word types, according to my own count. Like the *Ordinalia*, the meter is syllabic rather than rhythmic and each line contains seven syllables. Here is the first stanza:

Tays ha mab han speris sans
wy abys a levn golon
Re wronte ʒeugh gras ha whans
ʒe wolsowas y basconn
Ha ʒymmo gras ha skyans
ʒe ʒerevas par lauarow
may fo ʒe thu ʒe worthyans
ha sylwans ʒen enevow

Morton Nance (1949) points out that twenty three lines of *Pascon agan Arluth* are closely similar to lines in the *Ordinalia*. Compare:

“yn pub gwythres y coth thys
gorʒya ʒe ʒu hay hanow
ke ʒe ves, omscumunys
ʒe ʒyveyth veth yn tewolgow
ʒe vestry a vyth leʒes
neffre war an enevow”

Pascon Agan Arluth: stanza 17.

‘in everything that is done you should
worship your God and his name.
Go away, accursed one,
Into a wilderness so that you will be in darkness.
Your power will be diminished
Over the souls, for ever.

with

“yn pup maner y coth thys
gordhye the deu hay hanow
ke the ves ymskemenys
yn defyth yn tewolgow
the vestry a vyth leyhys
neffre war an enevow”

Passio Domini: lines 139-144.

‘In every way you should
worship your God and his name.
Go away, accursed one,
Into the wilderness, into darkness.
Your power will be diminished
Over the souls, for ever.’

Morton Nance (1949) concludes that there is neither agreement nor conclusive evidence concerning which way the borrowings went.

The earliest extant copy of the *Ordinalia* is in the Bodleian Library (Bodleian 791). It is described by Madan and Craster (1922: 405) as, “Bodl. 2639. In

Cornish, on parchment: … 11 × 7¾in., vii + 90 leaves …. Donum Jacobi Button armigeri ex comitatu Wigorniensi 28° Mart. 1615." The Bodleian Library's acquisition of this manuscript in 1615 is confirmed by Lhuyd (AB: 265) who also describes it as, "*Ex dono* Jacobi Button *Armigeri, è Comitatu* Wigoniensi. *An*. 1615." There appears to be no basis then for Hals' statement (British Library Add. MS 29762: f.90; also cited by Whitaker 1804: II, 24-5) that it was "brought into Oxford in 1450, and still extant in the Bodleian Library there". The manuscript (Bodleian 791) consists of 97 folios. After several blank pages, the first play begins on the eighth folio, numbered 1 in the top right hand corner. The next 82 folios, containing the three plays of the *Ordinalia* are similarly numbered. The folios have been written recto and verso, so the play is written on 166 pages or 83 folios in total. A few more blank pages follow folio 83. The whole of the *Ordinalia* contains 30,122 word tokens and 4,339 word types, according to my own count.

The *Ordinalia* is a cycle of three dramas. The first, *Origo Mundi* illustrates a number of Old Testament stories from the Creation to the building of Solomon's Temple. The second, *Passio Domini* illustrates the story of Christ's Passion. In "A Cornish Poem Restored", Morton Nance (1949: 368) asserts that a Cornish poem has been borrowed by the author of *Passio Domini*:

> It is into this Passion Play also that has been inserted, I think, a far shorter religious poem in which the Mater Dolorosa in beautiful Cornish verse makes what in the English of the time would be described as 'grete laymentacyoun'. In spite of being broken up so as to fit into three separate scenes of the play, none of it seems lost, and its singular metrical arrangement makes it easy to sort out and put it together again as forming two verses, each of twenty-five lines.

The third play, *Resurrexio Domini*, deals with the story of the Resurrection. The three plays were designed for open-air performance on consecutive days at parish feasts. The stage was a circular amphitheatre called in Cornish ‘plen an gwary’ or in English ‘playing place’.

The manuscript (Bodleian 791) contains the following graphemes: <A>, <B>, <C>, <D>, <E>, <F>, <G>, <H>, <I>, <J>, <K>, <L>, <M>, <N>, <O>, <P>, <Q>, <R>, <S>, <T>, <U>, <V>, <W>, <X>, <Y>, <Z>, and yogh <ȝ>. The graphemes <U> and <V> are homographic as are <I> and <J>.

A number of copies have been made of this original. There are two copies in the Bodleian Library. Bodleian Corn e 2 contains a transcript of Keigwyn’s 1695 translation of the *Ordinalia* by John Anstis the elder (died 1745). Bodleian Corn e 3 contains a copy of the *Ordinalia* revised and corrected by either Thomas Tonkin or William Hals. There are three copies in the National Library of Wales. NLW Peniarth 428 contains a copy by John Keigwyn of the *Ordinalia* with no translation. It bears the name Izabel Keigwyn on the first folio. Hawke (1979: 45) is of the opinion that the copy was made between the years 1695 and 1700 approximately. Davies (1939: 11) suggests that since it is from the Sebright collection it must have belonged to Edward Lhuyd. NLW Llanstephan 97 contains a copy of the *Ordinalia* with John Keigwyn’s autographed English translation and Latin preface. Hawke (1979: 45) is of the opinion that this copy was made circa 1702. According to Davies (1939: 8-10), this copy was made for Edward Lhuyd and was acquired by the National Library of Wales via John Williams’ library. NLW 21001 contains a transcript of the *Ordinalia* together with a transcript of Keigwyn’s translation.

According to Hawke (1979: 46) this was also formerly in the possession of Edward Lhuyd.

Concerning the place and time that the *Ordinalia* was written, Madan and Craster (1922: 405) describe the manuscript (Bodleian MS. 791) as "written in the first half of the fifteenth century in Cornwall (?)". Fowler (1961: 125) concludes that

> A re-examination of the place-name evidence suggests a date somewhere between 1300 and 1375, or more narrowly, between 1350 and 1375. …. The evidence of Middle English lines and phrases, vocabulary, and, above all, pronunciation of the final –e, point strongly to a date no later that 1400. … it is difficult to believe that the Middle English elements would allow a date earlier than the fourteenth century. …. It is possible to affirm, I believe, with some measure of confidence, that the evidence thus far considered points to the third quarter of the fourteenth century as the period in which to place the composition of the Cornish Ordinalia.

There is, however, good evidence for supposing that Bodleian 791 is of a much later date. Twice in Passio Christi, Jesus is referred to as the Son of Joseph the Smith. "Hemma yu an keth ihesu a lever y vos map deu map iosep an coth was gof" (*Passio Domini*: lines 1693-1695), 'This is the same Jesus who says he is the Son of God, Son of Joseph the old smith fellow'. "Cryeugh fast gans mur a grys may fo an ihesu crousys map an guas gof" (*Passio Domini*: lines 2477-2479), 'Cry aloud with much strength so that Jesus will be crucified, Son of the smith fellow'. These references to "an … gof", 'the smith', look very much as if they allude to Michael Joseph An Gof of St. Keverne, who was one of the leaders of the 1497 Cornish rebellion. When one considers that, following the rebellion, Michael Joseph An Gof was executed by the English, and that in *Passio Domini*, Christ's torturers speak phrases of

English, the case for this being an allusion to Michael Joseph An Gof appears even stronger. If this is the case, then Bodleian 791 may be dated circa 1500.

Several authors (Pedlar 1859: 504; Fowler 1961: 96; Bakere 1980: 12-49; Murdoch 1993: 41) have noted that place names mentioned in the *Ordinalia* suggest that it was written in the Penryn district. Pedlar (1859: 506) construes,

> If then we are to ascribe to an inhabitant of Penryn and to an ecclesiastic, the authorship of these plays, in as much as we find them written apparently shortly after the college of Glazeney was founded in that very place, we may conclude, with something like certainty, that they were the productions of that house.

Crawford (1980: 150), however, points out that it is only the place names attested in *Origo Mundi* that are clustered persuasively around Penryn. The two place names attested in Passio Christi are situated far to the West and those attested in *Resurrexio Domini* are dispersed over a wide area.

Sandercock (1984: 162) stresses that the oldest manuscript (Bodleian 791) is not the original. He points out that mistakes may well have taken place in the transposition which was written in two different hands, probably in the fifteenth century. Furthermore additions and alterations to this manuscript were made subsequently.

Edwin Norris (1859a) edited the *Ordinalia* with a translation into English. Norris (1859a: vi-ix) claims to have made his translation with the aid of Lhuyd's (AB) grammar and Pryce's (ACB) vocabulary and not to have seen the translation made by John Keigwyn (Bodleian Corn e 2). However, in my own personal copy of Norris (1859a: ix), there is a marginalis by E.G.R. Hooper, "unfair to John Keigwin – you gave up PRYCE and used Keigwyn:

E.G.RH". Stokes (CNACD) lists a number of corrections to Norris' (1859a) transcription. Loth (1905) also lists a number of corrections to Norris' (1895a) English translation. Morton Nance and Smith (Morton Nance n.d.) transcribed the *Ordinalia* into Unified spelling and produced a new English translation. Morton Nance and Smith's transcriptions and translations of *Passio Domini* (Sandercock 1982) and *Resurrexio Domini* (Sandercock 1984) were published by Kesva and Tavas Kernewek. Kesva and Tavas Kernewek also published the first 465 lines of Morton Nance and Smith's transcription and translation of *Origo Mundi* (Sandercock 1989). Morton Nance and Smith's transcription and translation of *Origo Mundi* was published in full in 2001 by Agan Tavas (Chubb, Jenkin and Sandercock 2001).

There have been a number of so called translations of the *Ordinalia*. Among them is a translation of *Origo Mundi* by Phyllis Pier Harris (1964) prepared from Bodley MS 791. There is also a translation by Markham Harris (1969) of the University of Washington, who claims:

> My goal was to produce, insofar as possible, a rendering that would prove responsible to the original and, what I think of as equally important from the literary point of view, responsive to the considerable range of tone to be found in Cornish.

Hooper (1972b) criticises the accuracy of some of these translations:

> `scholars' like the Americans who paraphrase Norris - Morton Nance – and - Smith and call it 'translation'. They wouldn't do that to Breton or Welsh - too many native speakers would have shot them down. But Cornish is safe for exploitation.

The manuscript of *Beunans Meriasek*, 'The Life of St. Meriasek', was discovered in 1869 by W.W.E. Wynne of Peniarth Library whilst he was

preparing the catalogue of the Hengwrt manuscripts (Williams 1869: 408). The manuscript (National Library of Wales MS. Peniarth 105) bears the colophon, “Finitur per dominum Rad Ton anno domini 1504”. “Rad” may be a shortened form of either ‘Richard’ or ‘Radulphus’, ‘Radulphus’ being the Latin form of ‘Ralph’. The entire play appears to be in Ton’s handwriting except for some corrections and stage directions made in another hand. The manuscript consists of 181 pages containing 21,010 word tokens and 4,401 word types, according to my own count.

The play is in verse throughout. Three plots are interwoven. The first concerns the life and death of St. Meriasek, who is associated with Camborne in Cornwall. The second plot concerns St. Sylvester the Pope and the Emperor Constantine. The third plot is that of a woman whose son is taken prisoner by a heathen tyrant and then miraculously released by the intercession of the Virgin Mary. It has been demonstrated how these three seemingly disparate plots are linked by the theme of “tyranny” (Payton 1993; Olson 1997). The subversive and political aspect of the play is noted by Jenner. Jenner (1928: 33). suggests that King Teudar, an evil tyrant depicted in the play, alludes to Henry VII, who was hated by the Cornish following the rebellion of 1497.

The manuscript (National Library of Wales MS. Peniarth 105) contains the following graphemes: <A>, <B>, <C>, <D>, <E>, <F>, <G>, <H>, <I>, <J>, <K>, <L>, <M>, <N>, <O>, <P>, <Q>, <R>, <S>, <T>, <U>, <V>, <W>, <X>, <Y>, long-tailed-z <ʒ>, yogh <ȝ>. The graphemes <U> and <V> are homographic as are <I> and <J>. The long-tailed-z grapheme <ʒ> represents dental fricatives and is attested in free variation with <TH>. The grapheme

<DH> is not attested in *Beunans Meriasek*. Yogh <ȝ> is homographic with long-tailed-z <ȝ> and is in free variation with <Y>. To confuse matters further, there are a few instances where <TH> is in free variation with <GH>. There is some free variation between <C> and <S>.

Williams produced a transcription of the first thirty-six lines which were published in *Archaeologia Cambrensis* (Williams 1869: 409). Whitley Stokes (1872) was the first to produce a critical edition of the entire play. Stokes (1872) edition includes his translation. Morton Nance and Smith (n.d.) made a transcription into Unified-spelling and translation into English entitled *Bewnans Meryasek*. Morton Nance and Smith's transcription has never been published in its entirety. However a few extracts have been published (Morton Nance and Smith 1966, 1969, 1974). Combellack-Harris (1985) completed a critical edition with English translation for a PhD thesis at Exeter University. Combellack-Harris (1988) also made a verse translation of the play.

The *Black Book of Merthen* is a survey of the Estate of John Reskymer of Merthen in the parish of Constantine. The survey was written between 1506 and 1536. On the title page, the Reskymer family arms are portrayed, incorporating the Cornish motto, "Keen awra". This motto may be translated as either 'I will make a reason' or 'I will do otherwise'. Within the text of the survey, the Cornish word "agomarocyon", 'knightly service', is found. This is the only known attestation of this Cornish word.

In his *Fyrst Boke of the Introduction of Knowledge*, Boorde (1555) aimed to aid the English traveller abroad:

> to teach a man to speak parte of all maner of languages, and to know the usage and fashion of all maner of countreys, and for to know the most parte of all maner of coynes of money the whych is currant in every region.

Apart from English, Boorde gives conversations in Lowland Scots, Cornish Dialect English, Cornish, Welsh, Irish, Low and High Dutch, Latin, Modern and Classical Greek, Ancient and Modern Hebrew, French, Italian, Castilian, Spanish, Turkish, Moorish, and Egyptian Arabic. Included in his "Apendex to the Fyrst Chapter, treatinge of Cornewall, and Cornyshe Men", are a smattering of "naughty English" (i.e. dialect English), Cornish "usage and fashion", the numerals in the Cornish language, and a conversation, in Cornish and English, between a traveller and the landlady and maid at an inn.

Sandys (1846) includes Boorde's section on Cornwall in his *Specimens of Cornish Provincial Dialect*. Zeuss (1853), Stokes (1879-1880) and Loth (1900) refer to Boorde's section on Cornwall in connection with Cornish numerals and conversation. Morton Nance (1928: 374-5) points out that Zeuss (1853 p. 325) has slipped into printing the English "eyght" instead of the Cornish 'eth', and that there are further errors of transcription to be found in Sandys (1846), Stokes (1879-1880) and Loth (1900). According to Morton Nance (1928: 375):

> The conversation gives little that is not to be found elsewhere, and its chief value lies in its phonetic spelling. This follows Boorde's system of writing the sounds of all foreign languages in the current English fashion. It is not very precise, and the Cornish dialogue is like the Welsh one in having its words split up or joined together without much regard to their meaning, ...

Morton Nance (1928 p. 381) is of the opinion that the "conversation is undoubtedly genuine". He suggests, however, that it might possibly have been

taken down “viva voce across a tavern table” from a Cornishman visiting London, in reply to Boorde’s questions such as “How do you say ... in Cornish?” In consequence it is possible that Boorde never visited Cornwall. Morton Nance (1929), in his *Cornish for All*, includes his Unified transcription of Boorde’s text along with the English.

Loth (1911b) describes a Cornish phrase found in a Star Chambers document dated 1547. It is not clear from Loth’s article, written in French, where the original document is held. Loth (1911b: 443) cites the document as “*Star Chambers* Henry VIII, 8/171-175” and says that he received it from the Rev. Taylor Vicar of Saint Just who in turn took the document from the “*Feudal Aids*”. According to this document, a certain John Richard of St. Just had a dispute with man called Carvanell who was the owner of a stamping mill for washing tin. The dispute concerned a crasing mill that Richard had built above Carvanell’s stamping mill. Carvanell’s farmer, Tracy, testified that he went one day to Richard’s mill in order to correct the rate of flow of the water. Surprised by Richard, he failed to turn on the water. Richard entered by the roof and evicted Tracy with blows, saying to him that if he ever found him in the vicinity of his mill again, he would not be held accountable for his actions. John Richard then went to Carvanell’s mill, and said to him in Cornish “deese meese te lader”, ‘come forth thou thief’. When Carvanell came out, Richard threw a large stone at him which fell between his legs. As he was trying to dodge, he was hit on the head by a shovel. He fell and Richard raised the shovel again. But the neighbours who were busy washing tin, ran up and restrained him. Tracy explained the problem for Carvanell’s mill, saying that

water is set from under the “polros” (‘wheel-pit of water mill’) of another mill. This word is clearly a compound of POL (‘pool’) and ROS (‘wheel’).

In 1555 a book was published entitled *A lyttle treatyse called the Image of Idlenesse conteynynge certeyne matters moued betwene Walter Wedlocke and Bawdin Bachelor. Translated out of the Troyane*[1] *or Cornyshe tounge by Olyuer Oldwanton and dedicated to the Lady Lust. Imprinted in London by William Seres dwellynge in Powles Churche yard at the signe of the Hedge hogge*. This book is a collection of bawdy tales of love and marriage written in English. The author claims to have translated the play from a Cornish original. In chapter seven, the following passage, containing a sentence in Cornish, occurs.

> Tyll at length this Pigmalion died and then was his wife turned agayne into an image of alabaster whiche to this day so remayneth and is accompted throughout all Greece theyr best and chiefest Pylgremage for to remove or expell the passions and paynes of ielousy.... The Princes of Tarent (but after some bookes, of Ottronto) ... being warned by a vision to repayre unto this blessed image for helpe, did avowe her Pylgramage thyther and receaved the Oracle, *Marsoyse thees duan Guisca ancorne Rog hatre arta* [my italics], being expounded by the prestes of that Temple to this effect in Englyshe. If to weare the horne thou fynde thy selfe agreeved. Gyve hym backe agayne and thou shalt sone be eased.
>
> (Oldwanton 1555: chapt. 7)

Jenner (1929: 239) is of the opinion that there was no Cornish original from which this book was translated. Berresford Ellis (1974: 68-69), however, sees no convincing grounds for Jenner’s assumption, and points out that bawdy Irish and Welsh tales indicate that this genre is to be found elsewhere in the

[1] Berresford Ellis (1974: 67) wrongly cites this word as “Troyance”.

Celtic literary tradition.

The so-called *Tregear Homilies* (Tregear n.d.) consist of John Tregear's translation of twelve of the homilies found in Bonner's (1555) *A Profitable and Necessarye Doctryne*… plus a thirteenth homily in a different hand and from an unidentified original. The entire manuscript consists of 130 pages containing 38,738 word tokens and 6,543 word types. It is thus the longest extant item of Cornish prose. The manuscript of the *Tregear Homilies* was discovered in 1949 by chance by John Mackechnie amongst papers of the Puleston family of Flint in Wales (Morton Nance 1950, 1951).

Tregear has been much criticised for the large number of loan words that he employs. Morton Nance (1951: 27) writes,

> … partly perhaps because he feared to tamper with the exact sense of the English, partly perhaps as falling in with the view of his congregation that high subjects needed loftier and less understandable language than popular Cornish, but chiefly I am afraid out of sheer slackness, Tregear when he came on a long English word like 'incomprehensibility', 'inclynacion', 'uncharitableness' etc. would very often leave it untranslated. He drags in simple English words, too, quite needlessly, such as 'even', 'only', 'not indeed worthy', 'by and by', 'meet', 'due', 'meek', 'lack', 'food', the Cornish of which he, of course, knew very well, and adverbs like 'chiefly', 'finally', 'wholly', 'freely', 'principally'. In one place he starts to write, in English, 'we have thereby' corrected immediately to 'us thynnu drethy': in another he corrected 'truth' to 'gwyryoneth' and in another writes only the capital T of 'take' before writing correctly 'kemereugh'. His clerical Cornish is in fact rather like the half French sermon language ridiculed in Brittany as Brezoneg Beleg 'Priests' Breton' ….

Among the *Exeter Consistory Court Depositions* for 1569-1572, in an entry for the year 1572, is a short phrase of Cornish (Hoblyn 1936: 11; Berresford Ellis 1974: 67). The entry runs thus,

> 1572. Wm. Fytteck of Lelant, tynner, resident from birth, aged 26, says he was in the parish church of Lelant on All Hallows day in the forenoon and at the time when the priest was at service, Agnes, wife of Moryce David and Cicely James came into the church, one after another, and were multiplying of words together and among their talk when that Cicely James was come almost to the mydle of the church, she called Agnes whore and whore bitch, and Agnes went in her pew and said nothing and there were a great many of the parish there that did hear the words. Wm Hawyshe, of Lelant, tynner, from birth resident, aged 40, sayeth that upon *dew whallan gwa metten in eglos de lalant* [my italics], viz. upon all hallow day late paste about the mydds of the service in the parish church of Lalant Moryshe David's wife and Cicely James came into the church of Lalant together and in chiding with words together Cycely called Agnes Davy whore and whore bitch in English and not in Cornowok.

Apart from the Cornish phrase "dew whallan gwa metten in eglos de lalant", it is significant that "whore" was said "in English and not in Cornowok", since Cycely James might have tried to claim, in her defence, that she had used the Cornish word, HOER, which means 'sister'.

Carew's (1602) *Survey of Cornwall* contains a few fragments of Cornish. These fragments include the phrase "Meea navidna cowzasawsneck" which Carew translates, 'I can speak no Saxonage'. In fact this phrase means 'I will not speak English'. Also amongst these fragments is the following prophecy: "Ewra teyre a war meane Merlyn Ara Lesky Pawle Pensanz ha Newlyn". This may be translated as, 'On Merlin's rock will land those who will burn Paul, Penzance and Newlyn'. Although Carew knew a few phrases of Cornish, he does not correctly segment these phrases into words. Thus he writes, "Molla tuenda laaz", which would be better segmented as 'Molla tue en da laaz' ('God's curse in your guts').

During the early eighteenth century, Tonkin produced an edition of Carew's *Survey*; however, this was not published until 1811. Halliday's (1953) edition

contains almost the whole of Carew's *Survey* as well as a long critical introduction.

The oldest extant version of the play *Gwreans an Bys* is in the Bodleian Library (Bodleian 219). It bears the colophon "Heare endeth the Creacōn of the worlde wth noyes flude wryten by William Jordan: the XIIth of August 1611". The manuscript consists of fifty four folios, containing 12,900 word tokens and 3,310 word types, according to my own count. The play itself, preceded by several blank pages, commences on the twenty fifth folio which is numbered "1" in the top right hand corner. The play is written recto and verso on the next twenty six folios, making a total of twenty seven folios or fifty three pages, the play ending on folio twenty seven recto. Folios twenty eight to thirty are blank. At the end of the play the audience are told, "dewh a vorowe a dermyn why a weall matters pur vras", 'Come tomorrow on-time; you will see very great matters'. This seems to indicate that the play is only the first part of a mystery cycle of which the remainder is missing.

The oldest of the extant manuscripts of *Gwreans and Bys*, Bodleian 219, contains the following graphemes: <A>, <B>, <C>, <D>, <E>, <F>, <G>, <H>, <I>, <J>, <K>, <L>, <M>, <N>, <O>, <P>, <Q>, <R>, <S>, <T>, <U>, <V>, <W>, <X>, <Y>, <Z>, long-tailed-z <ȝ>. The graphemes <U> and <V> are homographic as are <I> and <J>. The long-tailed-z grapheme, <ȝ>, represents dental fricatives and is attested in free variation with the graphemes <D> and <TH>.

According to Stokes (1863), the way the stage directions are set out and the

author's mention of limbo indicate that the play was written before the reformation. Morton Nance and Smith (1959) suggest 1530-1540 as the date of composition. In comparing the orthography with that of Boorde (1555), Morton Nance (1928) observes that Boorde's Colloquies display, "several late Cornish forms that are rarely or never found in Jordan's Creation of 1611 …." That Jordan's manuscript was transcribed from an earlier one has usually been assumed, but such forms suggest that his original might have been ten or twenty years older than the Boorde's colloquies, and thus nearly a century old in 1611. It seems unlikely, therefore, that Jordan was the original author. Possibly he was transcribing from an earlier manuscript, now lost. Whole passages are remarkably similar to *Origo Mundi*, and Neuss (1971: 129-137) suggests that *Gwreans an Bys* may have been constructed around the remembered part of one of the players who had taken part in *Origo Mundi*.

There is a manuscript copy of *Gwreans an Bys* in the British Library (Harleian 1867) and a copy of "The Creation, finished by J. Keygwin, gent. in ye year 1693" amongst the *Gwavas Manuscripts* (24v to 49r). There is a copy of *Gwreans and Bys* in Cornish with a copy of Keigwin's translation transcribed for the Bishop of Exeter in 1698 by John Moore (Royal Institution of Cornwall Cornish Play, The Creation 1698 - Common-Place Book of William Gwavas). There is an incomplete copy and incomplete translation in the hand of the Rev. Henry Ustick in the Bodleian Library (Bodleian Corn c 1). There is a copy of part of *Gwreans an Bys* in the Royal Institution of Cornwall (*Tonkin MSS B*).

Gilbert's (1827) edition of Keigwyn's transcription and English

translation has been much criticised for it's numerous inaccuracies (Norris 1859a: II 441; Stokes 1863: 1). Stokes (1863) produced an edition with a translation and notes. Morton Nance and Smith (1959) made a transcription into Unified spelling and translation into English which was edited for publication by Hooper (1985). There is an edition with translation into English by Neuss (1971). There is an English translation by Donald Rawe (1978).

Richard Brome's (1632) play, *The Northern Lasse: A Comoedie*, contains one sentence of Cornish. In Act V, scene 8, a character, disguised as a Spaniard, arrives. The other characters in the scene experience difficulty in conversing with him.

> Bullfinch: Alasse what shall wee doe then? Gentlemen have any of you any Spanish to help me understand this strange stranger?
>
> [They all disclaim knowledge.]
>
> Bullfinch: What shiere of our nation is next to Spain? Perhaps he may understand that shiere English.
>
> Tridewell: Devonshire or Cornwall, sire.
>
> Nonsense: Never credit me but I will spout some Cornish at him. Peden bras vidne whee bis creagas.

This sentence of Cornish may be translated as, 'Fat head, do you want to be hanged?'

In his *Archaeologia Britannica* (AB), Lhuyd includes three samples of Cornish text, all three written in Lhuyd's own phonetic notation. The first of these is Lhuyd's (AB: 222-224) own preface to his Cornish Grammar, "Dhan Tiz Hegaraz ha Pednzhivik Pou Kernou", 'To the Courteous and Noble inhabitants of Cornwall'. There is an English translation of Lhuyd's preface by

Thomas Tonkin (*Thomas Tonkin's MSS. B*) which was printed by Pryce (ACB). The second sample of Cornish is a short verse of three lines which Lhuyd (AB: 251) obtained from the Clerk of St. Just.

An lavar kôth yụ lavar guîr,
Bedh dɐrn rê ver, dhɐn tavaz rê hîr;
Mez dên heb davaz a gɐllaz i dîr.

This translates as follows, "The old saying is a true saying, a fist will be too short for a tongue that is too long; but a man without a tongue lost his land." The third sample of Cornish is the folk story of "Dzhûan tshei an hôr", 'John of Chyannor', which Lhuyd aligns paragraph by paragraph with its Welsh translation (AB: 251-253). This folk story consists of 46 paragraphs. Borlase (*Mems. of the Cornish Tongue*) made a transcription of "Dzhûan tshei an hôr" in his own adaptation of Lhuyd's General Alphabet. There is an incomplete version of this story in John Boson's hand (*Gwavas Manuscripts*: 128r-129r). There is also a transcription in Unified Cornish spelling by Morton Nance (1929: 38-48).

William Gwavas (born 1676 – died 1741), of Gwavas in the parish of Sithney near Helston, Cornwall, was a barrister and compiler of a collection of Cornish songs, verses, proverbs and letters (*Gwavas Manuscripts*). Gwavas's notebook (*Common-Place Book of William Gwavas*) includes some word lists, sayings and rhymes in Cornish, and a letter written by Gwavas and sent to John Boson.

The Gwavas collection of manuscripts (*Gwavas Manuscripts*) were formerly in the possession of Rev. William Veale of Trevaylor. Upon his decease they

passed to Rev. William Wriothesley Wingfield, Vicar of Gulval who presented them to the British Museum. The contents are from a variety of sources and in various hands. I list the contents here in full.

	DATE	PAGES
Letter from Davies Gilbert, Eastbourne, to Rev. W. Veale.	1836-07-22	1
Letter from John Boson, Newlyn, to W. Gwavas, Brick Court, Middle Temple, London.		2
Letter from W. Gwavas to Oliver Pendar, Merchant, Newlyn.		3
Letter from O. Pendar, Newlyn, to W. Gwavas, London.		4
“En levra coth po vo Tour Babel gwres” - elegy on the death of James Jenkin of Alverton, by John Boson.	1711/12-02-17	6r, 7r
Letter from W. Gwavas. Middle Temple, to John Boson, Newlyn.		8 to 9
Letter from John Boson, Newlyn, to W, Gwavas, Brick Court, Middle Temple, London.		10r
Letter from W. Gwavas to - - (aetate 55).	1731-03	11

Letter from John Boson, Newlyn, to W, Gwavas, Brick Court, Middle Temple, London.		12
Jottings for the translation of Matthew XIX, 17, by John Boson.		12v
Letter from Thomas Tonkin, Polgorran, to W. Gwavas, Penzance.	1735	14
Letter from W. Gwavas, Penzance, to T. Tonkin.	1735	16
Letter from Thomas Tonkin, Polgorran, to W. Gwavas, Penzance.	1735	18
Letter from W. Gwavas, Penzance, to T. Tonkin.	1735	20
Letter from Thomas Tonkin, Polgorran, to W. Gwavas, Penzance.	1735	22
Letter from W. Gwavas, Penzance, to T. Tonkin.	1735	23
Copy of “The Creation, finished by J. Keygwin, gent. in ye year 1693”.		24 to 29
The Lord’s Prayer in Cornish.		50
Copy of “Mount Calvary,” amended and corrected by W.H. (i.e. William Hals).	1679-1680	51 to 58

William Hals' "Lhadymer ay Kernou" A - Cluid (LK).		59ra to 78vc
Cornish Vocabulary - A to W.		80 to 89
Cornish Verses, &c.		91 to 97
The Ten Commandments in Cornish.		97 to 99
Genesis III in Cornish, by Wella Kerew.		99v to 101v
St. Matthew IV in Cornish, by Wella Kerew.		102r to 103v
St. Matthew II in Cornish, by Wella Kerew.		104r to 105v
The Creed in Cornish, by T. Boson.	1710	106
The Ten Commandments in Cornish, The Creed and The Lord's Prayer, by T. Boson.	1710	107 to 108
The Lord's Prayer.		109v to 110r
The Ten Commandments in Cornish.		110 to 114

Monumental inscription to be put on my tomb, William Gwavas, parish of Sithney, son and heir of Will Gwavas.	1719-09-16	166
Sundry Cornish writings, by W. Gwavas.	1731	167 to 168

The Gwavas collection of manuscripts (*Gwavas Manuscripts*) was transcribed in 1887 by John Gatley (1845-1936) of Trenewth in Michaelstowe. Gatley's transcription (Royal Institution of Cornwall, Gatley Transcript of Gwavas Collection) has several inaccuracies.

The Cornish historian Thomas Tonkin was born in 1678, in St. Agnes, Cornwall. Tonkin was a member of the lesser gentry who became Member of Parliament for Helston and participated in Stannary Court business. He was, like many of his contemporaries, involved in matters of tin and copper. Tonkin's general interests conformed to others of his social class of the time. He recorded his observations on antiquities, natural phenomena, and the general cultural environment in which he was immersed. Fortunately the Cornish language aroused his antiquarian interests.

Tonkin MSS B are amongst the manuscripts at the Courtney Library in the Royal Institution of Cornwall. The contents of this volume are described in Dr. Borlase's list of Tonkin manuscripts published in RIC Journal Vol. 6, page 168 (1879). The pages are all numbered in the top corner from 1-279. The Cornish language material is contained in Appendix "Numb. 1" (p. 171 ff.), Appendix "Numb. 2" (p. 205 ff.) and Appendix "Numb.3." (pp. 208 ff.). I

list all the Cornish language material here.

	PAGES
"A Cornish Vocabulary taken from the Originall in the Cotton Library, exactly as it is there written; only * the English is added to it for the benefit of the unlearned" * "And Modern Cornish."	171-192
Poem: "An Lavar Koth yw Lavar gwîr ...."	194
Poem: "In Obitum Regis Willielm ...."	195-197
"The Lord's Prayer, The Creed, And the Ten Commandment, In Cornish."	205-206
"The 1st Chapter of Genesis in Vulgar Cornish"	207-207.b
"Sentences in Vulgar Cornish"	207.b-207.c
"Proverbs"	207.c
Poem: "Ma Leiaz Gwreag ....", "Cornish Verses Composed for Curing Pilchards ...."	207 f
"A Cornish Song": "Pelea era why moaz, ...", "A Fisherman's Catch": A Mi a Moaz, ..."	207.g
"Names of Our Cornish Fields ..."	207.i

Tonkin MSS H, are also amongst the manuscripts at the Courtney Library in

the Royal Institution of Cornwall. They contain (p. 367) “the forme of words used in the Kernawish Tongue, in the Administration of the Sacrament,” by John Jackman Vicar of the parish of St Feock. Tonkin (*Tonkin MSS H*: 366) writes,

> John Jackman Vicar of this Parish, Aged 63 years, that dyed about 23 years Past, (Son to John Jackman Vicar of Kenwyn & Key) hath often declared to the Writer of those lines & many Others, that for many years after his Induction into this Vicarage, He was necessitated to administer the Sacrament in the Kernawish Tongue to the Aged People of the Parish, As his - Predecessors had done, because they did not understand the moderne Teutonick - or Mother Tongue to Us.”

There is also a collection of Tonkin’s manuscripts, usually known as the *Bilbao Manuscripts*, in the Biblioteca de la Diputación de la Provincia de Vizcaya, Bilbao, Spain. These Tonkin manuscripts were part of Prince Louis Lucien Bonaparte’s collection of papers on philological topics. When Bonaparte died in 1891, his wife left his papers to the libraries of Bilbao, San Sebastian and Pamplona. A photocopy of these manuscripts may be found in the Royal Institution of Cornwall. The pages of the photocopy are numbered in biro in the top right hand corner and enclosed in a circle. Jenner (1925) wrote a description of the contents of the *Bilbao Manuscripts*. I list the Cornish language items in the *Bilbao Manuscripts* here.

	PAGES
Letter from Wm. Gwavas to Thomas Tonkin, Penzance 27 ffeb: 1734, containing the Cornish sentences, “.... Ke, ker gen ol guz Krêvder, Dho ffiney. Go on with all your Strength to Finish. Tho ve guz Gwâz izal. I am your Humble Servant. Wm. Gwavas.”	38

Glossary (not alphabetical) of place-name elements; Signed Wm. Gwavas and dated Penzance 12 April 1735.	40-42
Notes concerning several Cornish words; Signed Wm. Gwavas and dated Penzance: 27: May 1735.	44-45
Undated letter from Wm. Gwavas; ".... mêr' a Gormola, Tha why, A wêth, thort, Gus Kâr Guîr. Wm. Gwavas."	63-64
Letter from Wm. G.. Newlyn 11 Decr. 1736, containing Cornish verses composed by John Boson.	
"The Lord's Prayer, Creed and 10 Commandments in Moderne and Antient Cornish."	77-80
Sayings in "Vulgar Cornish" and "with ye. English thereto".	81
Numbers in Cornish.	81
"Names of Fishes - in Cornish. with ye Etimology"; sentences, and verses in Cornish.	82.
"Advice from a friend in ye. Contry - to his Neighbour that went up to Receive 16000L in London"; ".... one Parsons Certificate to another to marry a Couple whose - Bans had been called . in ye Cornish Toung. Drake Proanter East, Tha Tobuy Trethell ...."; verses in Cornish.	83
"An Lhadymer ay Kernou" i.e. Tonkin's Dictionary (CLEV).	112-204

William Scawen (d. 1689) conducted research into and collected various fragments of the Cornish language. The *Enys Collection* in Cornwall Record Office, Truro, contains various papers in the hand of William Scawen. I list the Cornish language material included therein here.

	PAGES
Sayings: “Cows nebas Cows da ....” etc.	124b/1095
The Lord’s Prayer in Cornish, Welsh and Armoric	126a/1098
The Creed in Cornish, Welsh and Armoric	126a/1098

Thomas Tonkin made a copy of Scawen’s manuscripts (*Scawen Manuscripts*). An abridgement of Scawen’s manuscripts was published by Gilbert (1838).

William Borlase (born 1696 – died 1722) was Rector of Ludgvan. He made copies of a great many Cornish manuscripts. The Borlase collection of manuscripts (*Mems. of the Cornish Tongue*) is in the Cornwall Record Office (Cornwall Record Office DDEN 2000). The manuscripts are bound in a single volume but lack a single set of page numbers that run from start to finish of the volume. Most of the content of the Borlase collection of manuscripts can be found elsewhere. The list of contents below is prepared from a microfilm of the manuscripts (filmed the 27th January 1998). The numbers refer to the frames of that microfilm.

Notes for an Introduction to a projected treatise on the Cornish language. These notes were largely made use of in the preface to the Cornish vocabulary in *Antiquities Historical and Monumental of Cornwall*, published in 1754 and again in 1769.	007
"Mr Lhuyd's Cornish Grammar ... somewhat contracted ...." (AB) Table of contents.	008
"Literal Mem^{ms} of the Cornish Grammar, from Mr Edward Lhuyd, and other observations from the Cornish manuscripts." This is an abstract of the "Directions for reading Old British Manuscripts" in the Cornish grammar in Lhuyd's "Archaeologia Britannicum" (AB) and "Cornish Grammar Contracted".	009-028
"Rules and observations relating to the Cornish and other British dialects". Extracts from Lhuyd's "Archaeologia," (AB) Tit. I, "Comparative Etymology," and the preface to Tit. VIII, "A British Etymologicon."	029-033
"First Essay for an English-Cornish Vocabulary". Taken for the most part from Tit. VIII of Lhuyd (AB). The evident idea was to give Cornish words with the equivalent Welsh and Breton. Most of the Welsh words have been inserted, but the Cornish and Breton parts are very incomplete.	034-052

“PART II. Printed in ye End of the Antiquities. Cornish words digested under two Initials with their English. Ludgvan, 8 April, 1749.” This is the nucleus of what afterwards became the “Cornish - English vocabulary,” printed at the end of Borlase’s “Antiquities.” There are two alphabetical series of words, the second being a small supplement. It is evidently only a rough copy, and is not the copy from which the vocabulary was printed. Borlase’s pagination (at 1) begins again with this section.	054-175

"Exact copy of a fragment, viz: The beginning of a Cornish English vocabulary as left in MS. by the late William Gwavas Esq. ... The MS. Returned to Mr Veale of Trevailer, 1759". The copy begins, "An essay towards an Alphabetical Etimologicall Cornish Vocabulary with y^{e} signification thereof in English of the names of persons places Towns fields Tinworks & rivers &c: by Wm. Gwavas of Penzance Gen: Anno Dni 1738. A." Then follows a dedication "To my esteemed Friend the Reverend Mr. W^{m}. Borlase, Rector of Ludgvan in the County of Cornwall," in which the plan of the work is set forth, and mention is made of two existing manuscripts on the subject, "one by Mr. Tho: Tonkin of Polgorren, the other by Mr. Wm. Hals near Truro, called Cornish-English Vocabulary, lately in the hands of Mr. Tremain that married one of the daughters of Mr. Henry Hawkins an attorney at St. Austell, both of which I have perus'd." The vocabulary is a very short one, containing approximately 80 names.	176-178
"King Charles the first's Letter of Thanks to the Cornish" in English.	179
"King Charles the first's Letter of Thanks to the Cornish", "Translated (but not verbatim) into Cornish by the late M^{r}. John Keigwin". The original of this, containing, as copied here by Borlase, the English of the King's Letter, Keigwin's translation, and a letter from him to Gwavas, is in the *Gwavas Manuscripts*.	180

Sentences and Proverbs from Lhuyd's MSS.. These total 40 in all. A few of them were used in the Cornish Grammar in Lhuyd's *Archaeologia Britannica*, (AB) and a few of the proverbs were published by W. C. Borlase in the Journal of the Royal Institution of Cornwall in 1866.	211
A single four-lined stanza, beginning, "Proanter nei en Pleu Êst". This is a verse from a song composed by one John Tonkin of St. Just, a tailor. The whole song of seven stanzas is in the *Gwavas Manuscripts*. This verse is published in the Journal of the Royal Institution of Cornwall of 1866.	212
"Cornish from an old Romance of Mr. Boson's of Newlyn, called the Duchess of Cornwall's progress to the Land's End and & to the Mount" .	211-212
"Some Compositions in the Cornish Language". These consist of the following:	
(a) "The first chapter of Genesis in Cornish by the late Mr. Boson of Newlyn, from his own MSS.". This is the version which was printed at the end of Davies Gilbert's (1826) edition of the "Poem of the Passion", and with much revision at the end of Williams' (LCB) *Lexicon*.	213-215
(b) "On the Death of Mr. John Keigwin, written 20th of April, 1716, by Mr. Boson of Newlyn". In the *Gwavas Manuscripts* (93).	215

“Flo vye gennes en miz Merh ....”	228
“Chee Den krêv, leb es war Tyr ....”	228
“Hithow Gwrâ gen Skîans da ....”	228
“Cara, Gorthya, ha owna Dew ....”	228
“A Cornish Song from Mr. Tonkin’s MSS .... Pelea era why moaz ....” There is a slightly different version in the *Gwavas Manuscripts* in the handwriting of (and signed by) Edward Chirgwin.	229
“A fisherman’s Catch given by Capt: Noel Cater of St. Agnes to T. Tonkin Esq: 1698”. This can be found in *Tonkin MSS B* in the Royal Institution of Cornwall.	229
“The St. Levan Man of Tshei an Hor, from Mr. Lhuyd’s Archaeol. pa.251. The Translat. from T[onkin]’s MSS”.	230-233
“Cornish Names of Places in Hexameter verse, 1724” by the Rev. Mr. Whiting rector of Mawgon & St. Martins.	234
“Copy of the Cornish Vocabulary in the Cotton: Library London copy’d by the Revd Dr. Jer: Milles Chantor of the Church of Exeter, 1753”.	235-246
A letter from Jeremiah Milles to The Revd Mr Toupe, St. Martins, Cornwall.	247, 248

William Bodinar, a Mousehole fisherman, wrote a letter (*William Bodinar’s*

Letter), in Cornish and English, to Daines Barrington. Barrington subsequently published the letter in *Archaeologia - the Journal of the Society of Antiquities* (Barrington 1776). William Bodinar's original letter (*William Bodinar's Letter*) is in the possession of the Society of Antiquaries. Here is the letter in its original spelling as published by Barrington:

> Bluth vee Ewe try Egence ha pemp
> my age is three score and five
> Theatra vee dean broadjack an poscas
> I am a poor fisher man
> me rig deskey Cornoack termen me vee maw
> I learnt Cornish when I was a boy
> me vee demore gen cara vee a pemp dean moy en cock
> I have been to sea with my father and five other men on the boat
> me rig scantlower clowes eden ger Sowsnack cowes en cock
> and have not heard one word of English spoke in the boat
> rag sythen ware bar.
> for a week together
> no rig a vee biscath gwellas lever Cornoack.
> I never saw a Cornish book
> me deskey Cornoack mous da mor gen tees coath
> I learnd Cornish going to sea with old men
> na ges moye vel pager pe pemp endreav nye
> there is not more then four or fiue in our town
> ell clappia Cornish leben
> can talk Cornish now
> poble coath pager egence blouth.

old people four score years old

Cornoack ewe all neceaves gen poble younk

Cornish is all forgot with young people

There is an article about the letter, by Pool and Padel (1975-1976), with full text, commentary and a facsimile of the letter.

There is a short phrase of Cornish attributed to Dolly Pentreath (born 1692 – died 1777) of Mousehole, née Doaryte Pentreath. In his book *Traditions and Hearthside Stories of West Cornwall*, Bottrell (1870: 184) tells the story of Dolly Pentreath's encounter with Mr Price of Choone, in which she calls Price, "Cronnack an hagar dhu", 'You ugly toad!'. Bottrell (1870: 185) gives his source for this anecdote as "an old lady of Sennen, who knew Dolly well".

Polwhele (1816: 43) cites an epitaph for Dolly Pentreath written by a mining engineer from Truro called Tompson. According to Polwhele, who met Tompson in 1789, Tompson knew more Cornish than Dolly had ever done. I quote the epitaph in full.

Coth Doll Pentreath cans ha deau

Marrow ha kledyz ed Paul pleû

Na ed en Egloz, gan pobel braz

Bes ed Egloz-hay, coth Dolly es.

This is translated as follows.

Old Dolly Pentreath, aged 102,

Deceased and buried in Paul parish too,

Not in the Church with people great and high,

But in the churchyard doth old Dolly lie.

J. Hobson Matthews (1892) cites a poem by John Davey (born 1812 – died 1891) of Boswednack near Zennor.

A Granken, a granken,
a mean ow gowaz o vean
ondez Parc an Venton
pub trelowza vean
Far Penzans a Maragow
Githack Macrow
a mac trelowza varrack.

Morton Nance (1922-1925) gives a respelt transcription and English translation of Davey's verse. Weatherhill (CPNL: 148) gives a transcription in Modern Cornish and the following English translation.

O Crankan, O Crankan
On stone one finds but little
Beyond the Well Field
That bears 3 shoots per stone
The Penzance to Marazion road
Both greatly more green
And greatly more fresh
Grows 3 shoots per horseman.

Figure 13 shows the comparative size in word types and word tokens of the five largest texts found in the Corpus of Cornish. It can be seen that the *Tregear Homilies* (Tregear n.d.) is the largest of these texts, followed by the *Ordinalia*. Together these five texts amount to 112,861 word tokens.

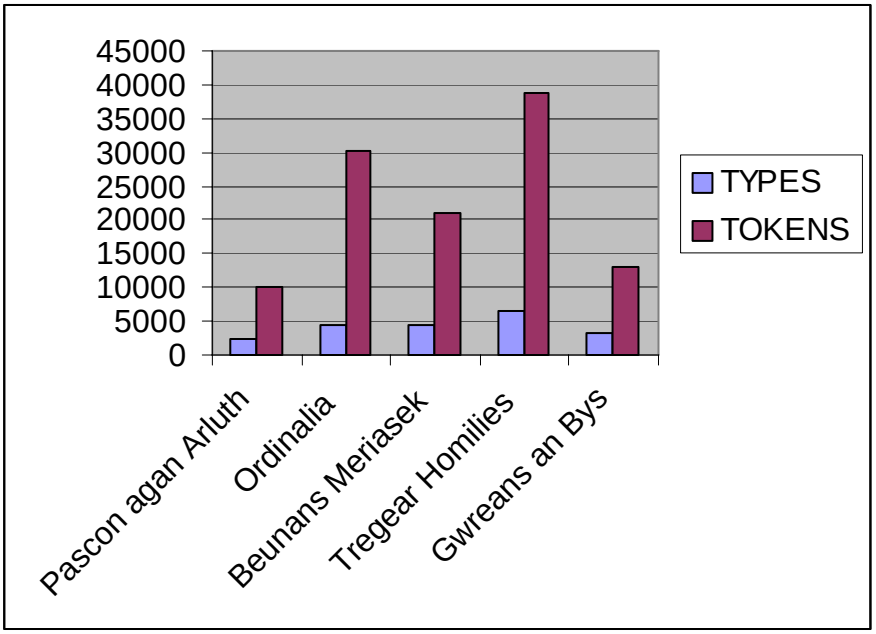


Figure 13 Comparative size of the main corpus texts

3.2 Methodology for compiling a historical corpus

The creation of new critical editions of the source texts for the Corpus of Cornish was undertaken. The corpus was initially prepared on computer as raw ASCII text files with no mark up of any kind. Wherever possible the oldest extant manuscript served as the source. The published transcriptions of Norris (1859a) and Stokes (1861, 1863, 1872) were a help in reading the original source manuscripts. However it was felt that Norris' and Stokes' published transcriptions contained too many errors to be used as they were. A fundamental concern regarding the compilation of the corpus was the methodology of tokenisation. The need for a systemic approach to tokenisation for lexicographical purposes was highlighted by the lack of correspondence between orthographic words and morphosyntactic words. The conversion of handwritten manuscripts into electronic tokenised critical

editions was made possible by a theoretical framework that incorporates the notions of token, type and tone. The lexical item is the unit of tokenisation for lexicographical purposes. Lexical items are located on a scale of rank as either morphemes, words or multi-word lexemes. It is not a straightforward matter to identify multi-word lexemes for tokenisation purposes.

The first stage of lemmatisation involves the segmentation of the corpus into tokens. During this stage, outer selection takes place, in which lemma signs are selected from the corpus (cf. Wiegand 1984: 596 ff.). 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. Some languages such as Chinese and Japanese, however, do not have explicit token delimiters. In such languages a sentence is a string of characters with no English blank space equivalent. Languages with conjunctive orthography such as Finnish and Shona have few word delimiters. Such orthographic practices make tokenisation a serious problem for Natural Language Processing in these languages. The problem also exists when working from the medieval manuscripts which are the source of the Corpus of Cornish, because words are not clearly delimited by spaces.

Sentence word tokenisation is the process of converting a sentence into a string of words. Because most Natural Language Processing applications take words as basic processing units, it is a common stage in the preprocessing of a text. Given a string of characters generated by removing blank spaces that function as word delimiters from a natural language sentence, a natural problem is to discover a way to restore these blank spaces (Guo Jin

1996: 1).

Leech (1997: 21-24) identifies three ways in which orthographic word tokens fail to correspond to morphosyntactic word tokens: multi-words, mergers and compounds. In the case of multi-words, a single morphosyntactic word corresponds to more than one orthographic word. Some multi-words are discontinuous. When a single orthographic word token corresponds to two or more morphosyntactic words, Leech calls this a "merger". In the case of compounds, one or several orthographic words, depending on the analysis, corresponds to one or several morphosyntactic words. Leech admits that this is a rather open-ended category and that a grey area exists between analysis as a single compound or as a sequence of two stand-alone nouns.

All three ways in which orthographic word tokens fail to correspond to morphosyntactic word tokens can be found in Cornish. "Tâz gwidn", 'a grandfather' (AB: 3) is an example of compounding. An example of a Cornish multi-word expression is the interjection, "gwyn ow bys" (*Gwreans an Bys*: line 2005), consisting of "gwyn" ('fair'), "ow" ('my'), and "bys" ('world'). "Gwyn ow bys" can be roughly translated as 'lucky me!'

In some languages the boundary of the word is not always clear. Thus the lexicographer will encounter items which it is difficult to classify as words or as morphemes. An example of such a merger is "han" (*Gwreans an Bys*: line 293), which maybe decomposed into the conjunction HA**,** 'and', and the definite article AN, 'the'.

Mergers frequently involve clitic forms. Zgusta (1971: 241) suggests that

these, too, should have their own entries, and that grammatical items, such as clitics, should not be overlooked simply because they seem to be of lower status than other words. A clitic may be defined as a sort of obligatory bound morph which is generally distinguished from an affix. A clitic may be a reduced form of a word such as English '-'ve' for 'have'. Both clitics and affixes are bound morphs. A distinction is drawn, however, between affixes, which are inflectional or derivational, and clitics, which are not. Furthermore affixes are usually attached to particular lexical categories. A clitic, on the other hand, is attached to a phrasal group or a single word in that phrase. A clitic may be attached to various parts of speech and can be attached freely to other affixes or other clitics. Clitics are usually divided into two categories, proclitics which are attached before their base, and enclitics which are attached after their base (Bauer 1988: 99-100). Thus the item "nynges" (*Gwreans an Bys*: line 426) may be analysed as consisting of the proclitic negative particle NYNG, 'not', followed by *es*, 3rd person singular, present tense of the verb BOS, 'be'. Similarly the item "theth" (*Gwreans an Bys*: line 629) may be analysed as consisting of the preposition THE followed by the enclitic pronoun ATH.

The lexical item is a minimal semantic unit with an identifiable form and is not necessarily directly related to specifically morphological units, words or phrases. The form of the lexical unit can in fact be any of these (Newell: 46). The way that a corpus is segmented into lexical items depends on the way that they are to be represented in the dictionary.

The token-type distinction originates with the philosopher Charles Peirce

(1931-1958: 4.1, 4.537, 2.245). 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. 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. 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, usually alphanumeric characters are considered unanalysable. In other words, the reader does not analyse them into bars and curves, or distinguish them according to pitch in proportional spacing.

Figure 14 shows an extract from the 16th century Cornish miracle play, *Beunans Meriasek* (f. 19).

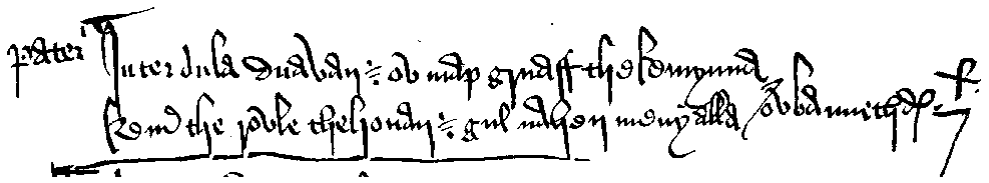


Figure 14 Extract from *Beunans Meriasek*

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 manuscript 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 15 shows the tone of the first occurrence of *the*.

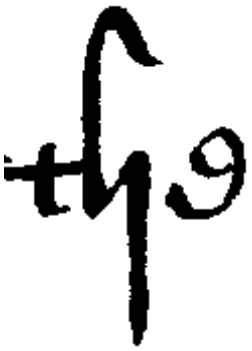


Figure 15 First occurrence of *the*

In order to read the original manuscript, it is necessary firstly to define the alphabet that is employed, and secondly to segment text into lines and lexical items. Lexical items may be selected at the rank of word, the rank of morpheme or the rank of multi-word lexeme.

In scale-and-category theory (Halliday 1961), grammatical form is concerned with the nature of elements of structure and relationships which may be established between them. The category of unit is concerned with stretches of language of varying lengths and composition. These units carry or operate in grammatical patterns and are related hierarchically at the scale of rank (see Figure 16). The sentence is the largest grammatical unit whilst the morpheme is the smallest. Each unit, except that of sentence, is defined by its function in the structure of the rank above. Conversely each unit, except that of morpheme, is comprised of one or more units of the rank below.

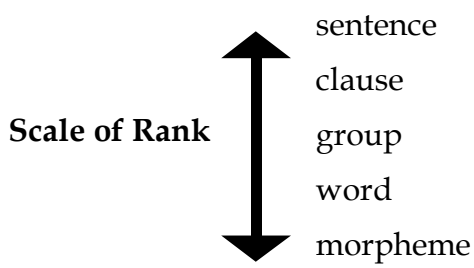


Figure 16 The scale of rank

During lemmatisation, the lexicographer is concerned with essentially three units at the scale of rank (see Figure 17). Two of these, word and morpheme, are to be found in Halliday's (1961) units. The third, that of multi-word lexeme, is specific to lemmatisation and may be realised by several types.

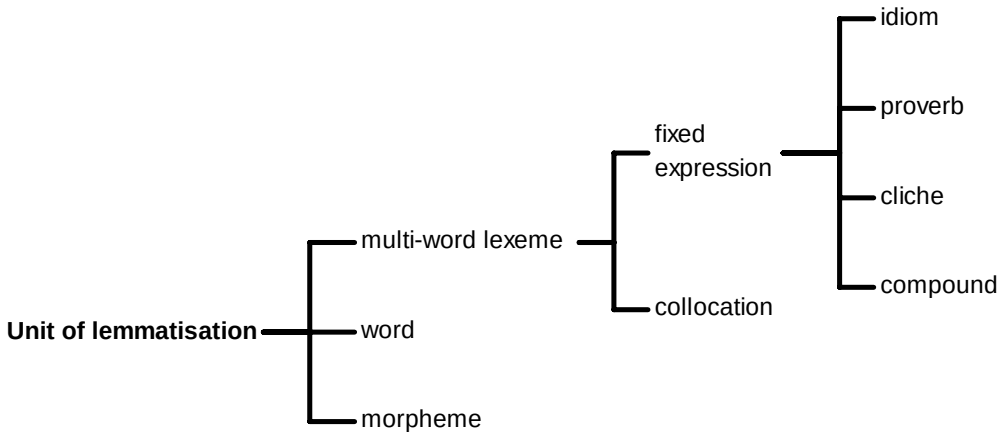


Figure 17 The unit of lemmatisation system

Let us first consider the rank of morpheme. According to Zgusta (1971: 157-8),

> … we may conceive language as consisting basically of units of two sorts, distinctive and meaningful. The distinctive units are the phonemes; they do not interest the lexicographer too much per se. The minimal meaningful units are the morphemes. Morphemes, however, are not directly constituent parts of the sentence: this function is performed by the lexical units, a very frequent morphological category of which are the words.

A morpheme is considered to be the smallest unit of meaning and each morpheme has its own meaning. Morphemes are not usually easy for the dictionary user to identify. So, for the purpose of a practical dictionary, the word is the unit of rank normally used. Zgusta (1971: 241) recommends that morphemes of highly productive prefixes and compositional elements need to have entries in the dictionary.

Morphemes fall into two categories, free and bound. Free morphemes can stand on their own and cannot be subdivided into smaller morphemes. Free morphemes always consist of a root. A bound morpheme cannot stand on its own and has to be connected to another morpheme (Burgess 1964; Nida 1976).

Next let us consider the rank of word. The word, as represented by orthographic tradition, is the most usual lexical unit. Words are traditionally bounded by spaces and are, thus, easily identifiable. Morphemes are not so easy for the dictionary user to recognise. As a result, the word is the grammatical unit best suited to the purpose of a practical dictionary (Mathiot 1967; Zgusta 1971: 240). According to Swanson (1975: 64),

> Current linguistic analysis provides us with more precise (though not yet definitive) criteria for "words." For IE languages in general there is no important problem either in the nature of the word or of its categories. In a language like English, phonetic criteria may determine the words (usually but not always coinciding with spelled words); for "archaic" IE languages morphology will have to suffice as our clue.

Finally let us consider the rank of multi-word lexeme. Certain expressions consist of several distinguishable lemmata (Schnorr 1991: 2815). Zgusta (1971: 241) recommends that multi-word lexemes should be selected, treated, and indicated as wholes, since they are of the same standing and function as single words and they are, in fact, treated in most dictionaries as a single entry word. The phenomenon according to which a string of several words is used to express a notion that is not analysable and distributable over the different words of the string is referred to as idiomaticity (Béjoint 1994: 210). Cowie (1981: 233) maintains that "lexical units are *complexes* of various kinds more often than the traditional organisation of the dictionary has prepared us to believe or reductionist images of the lexicon encourage us to suppose." According to McCarthy (1988: 56), "much of language comes in pre-packaged strings which display a limited number of patterns, as opposed to ... the classical linguistic notion that language consists of a series of syntactic 'slots' into which lexical items may be deposited."

Rey-Debove (1971: 113) asks why it should be that some syntagms are included as entry forms in dictionaries as opposed to other syntagms. Zgusta (1971: 144-52) distinguishes nine criteria which may be employed to detect multi-word lexemes.

1. Substitution is not possible in a multi-word lexeme.

2. It may not be possible to add something to the multi-word lexeme.

3. The meaning of the whole combination is not fully derivable from that of its single parts.

4. A constituent part of a multi-word lexeme may be severely or exclusively restricted to it.

5. The multi-word lexeme may have a one-word synonym or a close near-synonym.

6. Analogous or identical status among the multi-word lexemes and single words may be exhibited by a small group of semantically related expressions.

7. A one-word translation equivalent in a foreign language may indicate a multi-word lexeme. Osselton (1995: 99) points out that phrasal verbs that occur in the 'Promptorium Parvulorum' (15th century dictionary with an English alphabetical list) "are often there because they correspond to single Latin lexemes: original Latin-English entries have become reversed to give English-Latin ones". Translation equivalence was clearly a criterion used by Morton Nance (NCED: iii) who writes, "Hyphens are used to link words that are translated by one word in English ...."

8. Formal and grammatical properties are sometimes inherent in multi-word lexemes.

9. A multi-word lexeme performs syntagmatically in a

> sentence and paradigmatically in the lexicon the same syntactic and onomasiological function as a morphologically more simple unit which frequently coincides with the word. This fundamental requirement is the criterion by means of which set combinations of words like proverbs, sayings, dicta, quotations, and similar fossile, petrified expressions are distinguished from the multi-word lexemes.

One type of multi-word lexeme is the fixed expression. Fixed expressions include idioms, proverbs, similes and clichés. Examples of Cornish similes include “maga fery avel hok” (*Beunans Meriasek*: line 1901), ‘as merry as a kite’; and “maga whyn avel an leth” (*Passio Domini*: line 3138); ‘as white as the milk’. There also a number of Cornish proverbs and maxims such as “Nyn ges goon heb lagas na kei heb scovern”, ‘There’s no down without eye nor hedge without ears’ (*Enys Collection*: 1095); and “Na reys gara an vor goth rag an vor noweth”, ‘Do not leave the old road for the new road’ (*Enys Collection*: 1095). According to Zgusta (1971: 153),

> The lexicographer’s interest in these quotations, dicta and proverbs is rather negative. They are certainly set combinations of words, they must be understood as wholes, but they are not multi-word lexical units. Knowledge of them undubitably belongs to the knowledge of the language (and even more to knowledge of the respective culture), so that really big dictionaries may register them; but it should never be forgetton that though they are set groups of words, though they are understood as wholes and are frequently presented in the texts as wholes (intonation; quotation marks etc.), they are not single lexical units: they are built up of several of these lexical units in each case.

Another type of multi-word lexeme is the collocation. Hausmann (1989: 1010) defines a collocation as a string of words that is recognised by the language user as a ‘normal’ construction and which is bound together by its syntactic

structure. Binary collocations, such as N+Adj, V+Adj and Adj+Adv, are typical. However collocations of three or more words are also possible. Usually a collocation is a sentence constituent. Gorcy distinguishes between legitimate collocations that are *langue* and require proper attention, and those that are stylistic or *parole*. Lexicalisation refers to the process of a free syntagm become gradually frozen and this too must be accounted for (Gorcy 1989: 909).

Typicality may also provide a criterion for decisions about whether to treat a word string as a single lexical item. According to Béjoint (1994: 210), "Each element of language has typical uses which can be described in terms of syntactic and semantic environments, the whole thing being captured in a statistical study of text in terms of frequency". Zgusta (1971: 151), however, is sceptical of statistics as a means for identifying multi-word lexemes:

> I do not know of any conclusive count which could give us some undubitable examples. ... And I strongly suspect that the frequency of the co-occurrence of two words may be even greater if we have to do with a fully free combination of words which have themselves a high frequency of occurrences; e.g., a statistical count would probably show that the combination to drink beer has an immensely higher frequency of occurrence than to swallow stones, but both are free combinations anyhow.

There are essentially two approaches to word tokenisation, character based tokenisation and lexicon based tokenisation. Each of these algorithms has its own advantages and disadvantages.

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.

Figure 18 illustrates an algorithm for character based tokenisation. Various writers give Prolog code for character based tokenisers (Clocksin & Mellish 1981: 87-8; O'Keefe 1990: 319-54; Gal *et al.*1991: 232-3; Covington 1994: 318-20).

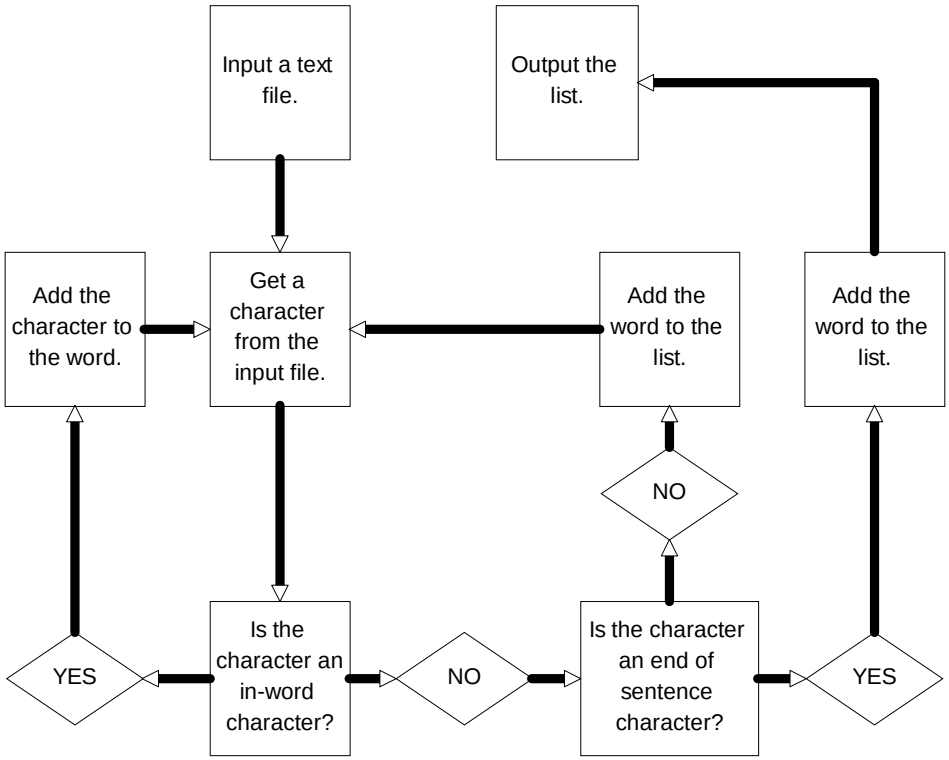


Figure 18 Algorithm for character based tokenisation

There can be, however, certain problems with this approach. In a language such as English, for example, some characters occur both within and between words. The hyphen occurs both within compounds and between parts of a sentence. The <'> symbol is used both within words as an apostrophe and as a closing inverted comma. The space can occur within a single lexeme such as *New York*. The comma can occur within a number such as *1,000*. The full stop

can occur within a time such as *12.45 pm*. Conversely, in English, words are not always delimited by characters, for example *gimme* and *haven't*. Character based tokenisation does not provide a foolproof way of segmenting text into tokens. It does, however, act as a guide to the eye when reading a text. Character based tokenisation may be fine tuned by employing a lexicon of exceptions. This approach is a combination of character based tokenisation and lexicon based tokenisation. Whilst character based tokenisation works fairly well for some languages such as English, other languages, such as Chinese and Japanese do not have explicit word delimiters. In the extract from *Beunans Meriasek* in Figure 14, 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 19 illustrates an algorithm for lexicon based tokenisation.

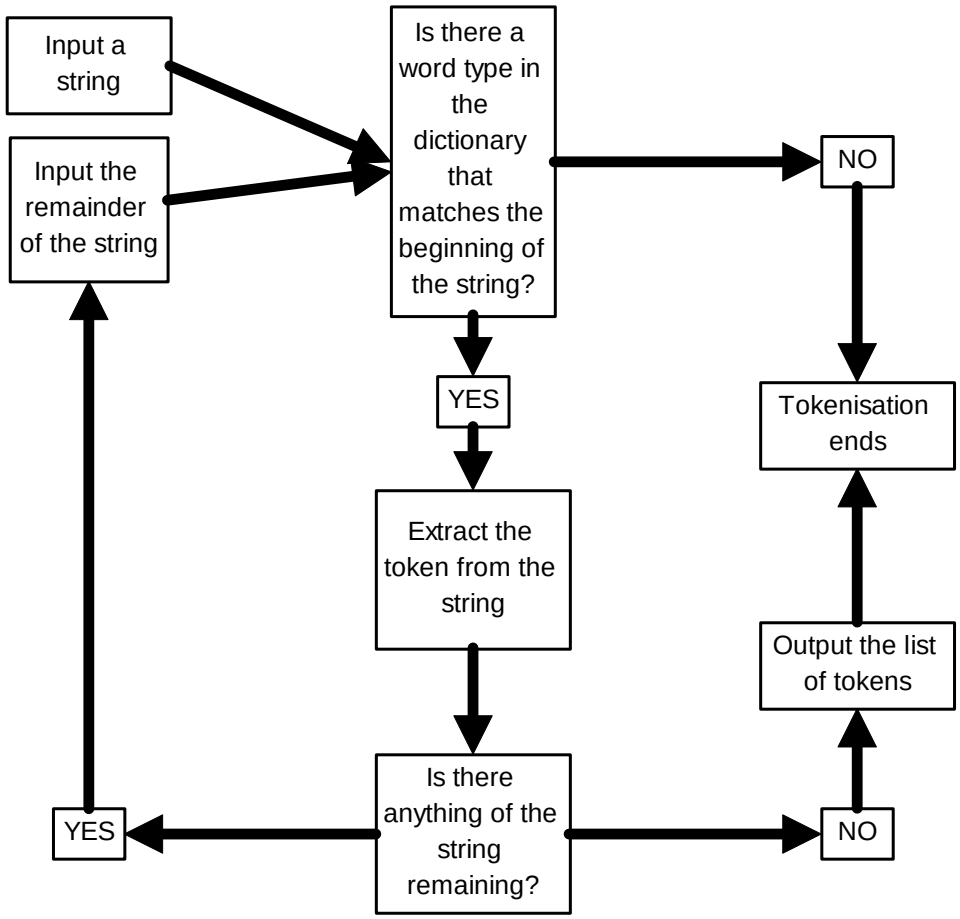


Figure 19 Algorithm for lexicon based tokenisation

Figure 20 represents a very simple dictionary to be used with the algorithm in Figure 19.

Dictionary

```
a
bird
black
blackbird
peter
saw
```

Figure 20 Simple dictionary for lexicon based tokenisation

Given the algorithm in Figure 19 and the dictionary in Figure 20, if we input

the string,

“petersawablackbird”,

the system will return the output,

[peter,saw,a,black,bird].

There are two problems with this algorithm. Firstly, if the string contains items which are not in the lexicon, tokenisation fails. Thus the string,

“marysawablackbird”,

will not tokenise. Secondly, a given critical segment may contain one, or more than one, word type. Alternative readings are, thus, possible. The algorithm, however, finds only one solution. The order in which items are listed in the dictionary determines the outcome. So since *bird* and *black* are both listed in the dictionary before *blackbird*, the system selects [black,bird] and [blackbird] is not chosen.

There are two types of ambiguity involved, combinatorial ambiguity and overlapping ambiguity (Guo Jin 1996). Combinatorial ambiguity refers to critical segments that consist of one or more than one word type. Figure 21 shows some examples of combinatorial ambiguity.

blackbird	*black bird*
below	*be low*
today	*to day*

Figure 21 Examples of combinatorial ambiguity

Overlapping ambiguity may be defined as follows. Given a character string

ABC, if the sub strings *A*, *AB*, *BC* and *C* are all words in dictionary, the string *ABC* is said to have overlapping ambiguity, as there exists an overlap between the word *AB* and the word *BC*. Thus, in English, the string “fundsand” can be tokenised as either “funds and” or “fund sand”. Similarly the string “toplace” can be tokenised as either “to place” or “top lace” (Guo Jin 1996: 3).

The algorithm in Figure 19 was implemented in the Prolog programming language. Prolog is a declarative programming language which uses unification to find all the solutions to a goal. If we run our Figure 19 algorithm in Prolog by setting the goal

?- tokenise(“petersawablackbird”,X).

then the system will find all the possible tokenisations of the string, thus

X = [peter,saw,a,black,bird];

X = [peter,saw,a,blackbird].

However, tokenisation will still fail if the string contains a word that is not in the dictionary. A complete dictionary is, therefore, needed. A complete dictionary is one in which all valid words are included and there are no unknown words. Guo Jin (1996: 2-3) points out that although a linguistically complete dictionary is never within reach, an operationally complete dictionary is trivial to compile. One simple way is to add all the characters in the alphabet to the dictionary as single character words. These then spell out unknown words which can be glued back at a later stage. Guo Jin’s solution thus combines lexicon based tokenisation and character based tokenisation.

```
?- tokenize(“petersawablackbird”).
[peter,saw,a,blackbird]
[peter,saw,a,black,bird]
[peter,saw,a,black,b,i,r,d]
[peter,saw,a,b,l,a,c,k,bird]
[peter,saw,a,b,l,a,c,k,b,i,r,d]
[peter,s,a,w,a,blackbird]
[peter,s,a,w,a,black,bird]
[peter,s,a,w,a,black,b,i,r,d]
[peter,s,a,w,a,b,l,a,c,k,bird]
[peter,s,a,w,a,b,l,a,c,k,b,i,r,d]
[p,e,t,e,r,saw,a,blackbird]
[p,e,t,e,r,saw,a,black,bird]
[p,e,t,e,r,saw,a,black,b,i,r,d]
[p,e,t,e,r,saw,a,b,l,a,c,k,bird]
[p,e,t,e,r,saw,a,b,l,a,c,k,b,i,r,d]
[p,e,t,e,r,s,a,w,a,blackbird]
[p,e,t,e,r,s,a,w,a,black,bird]
[p,e,t,e,r,s,a,w,a,black,b,i,r,d]
[p,e,t,e,r,s,a,w,a,b,l,a,c,k,bird]
```

Figure 22 Possible solutions of lexicon based tokenisation

Figure 22 shows all the possible solutions to the string when all the characters in the alphabet have been added to the dictionary as single character words. The solutions have been sorted so that the solutions with the least number of tokens occur at the top of the list. **`blackbird`** is, thus, listed before **`black,bird`**, and **`black,bird`** is listed before **`black,b,i,r,d`**. This manner of sorting I have named ‘Longest First Tokenisation’. Its purpose is to present what are possibly the more plausible solutions first.

Critical tokenisation is a way to represent a text that has been tokenised in such a way that both individual tokens and individual types can be identified. Lager (1995: 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 23 shows how critical tokenisation can be applied to the first line of the fragment of the Cornish miracle play shown in Figure 14.

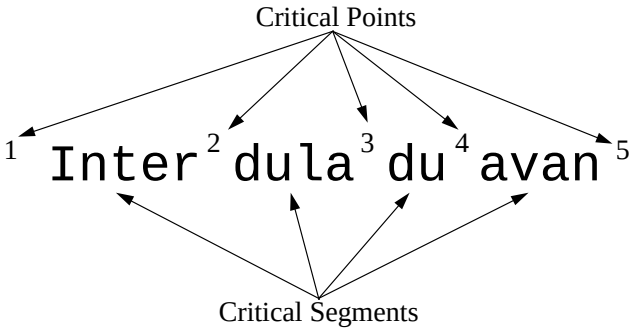


Figure 23 Critical tokenisation

Figure 24 shows how the principles of critical tokenisation may be applied to the fragment of the Cornish miracle play shown in Figure 14, in order to create a Prolog database. The text is arranged vertically with each token on a separate line. 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 24 Critical tokenisation implemented in Prolog database

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 word lists, concordances, lists of collocations and other types of data.

Of the two algorithms for corpus lemmatisation under consideration, it is noted that, since it is based as it is on the orthographic word, character based tokenisation does not cope well with the three ranks at which lexical items occur. In contrast, lexicon based tokenisation is able to recognize items that are realised at different points on the scale of rank. In addition, instances of combinatorial and overlapping ambiguity are identified by lexicon based tokenisation. Critical tokenisation as Prolog files provided the means to represent the tokenised texts. It is possible to manipulate the resulting text database in Prolog to retrieve types and tokens, and to produce word lists and concordances.

4 The Lemma

The lemma may be considered from two perspectives: according to the available literature on lexicographical theory, and according to the history of Cornish lexicography. A principal concern of lemmatisation is to unite the variant forms of the lexeme under a single canonical form. The wide variation in which Cornish lexical items are attested is either synchronic or diachronic.

There have been a number of attempts to define the term lemma. Landau (1989: 319) notes that the term lemma is sometimes used rather loosely to signify any word or phrase glossed or defined. Zgusta (1971: 249-251) uses lemma to refer to both the canonical form and its pronunciation. Landau (1989: 319) writes that he prefers to distinguish between a head word and its pronunciation, and, therefore, avoids using the term lemma at all. Hartmann & James (1998) define the lemma as the "position at which an entry can be located and found in the structure of a reference work".

Ilson (1988: 73) distinguishes the term lemma from head word and entry word by proposing that it is extended to mean 'everything preceding the first explanation (or sense number) in a dictionary entry'. Zgusta (1971: 249 ff.) maintains that, whilst the head word is the most important part of the lemma, it also necessary to include part-of-speech, pronunciation and sometimes etymology. The lemma may thus be described as the first of 2 parts of the entry, indicating the lexical unit itself. The purpose of the lemma is firstly to identify the lexical unit, secondly to locate it in the morphological system, and thirdly to describe its form, which may include indications of pronunciation.

The alphabetised head word thus represents a paradigm (cf. Zgusta 1971: 249ff.; Landau 1989: 76).

For the purposes of this project, the following definition is used. The lemma is a code the purpose of which is firstly to identify and unambiguously distinguish the lexical item from all other lexical items in the dictionary, and secondly to determine the position of the lexical item in the macrostructure of the dictionary. The form of the lemma typically consists of a number of fields which serve to identify the lexical item. These may include the canonical form, pronunciation, part-of-speech, genre label and various other fields. The lexical item that is represented by the lemma, may be any item which the lexicographer chooses to include for entry in the dictionary, including words, derivatives and compounds, word forms, flectional morphemes, affixes, affixoids, further elements of word formation, radicals, multi-word lexical units, phrasal verbs, parts of multi-word lexical units without monemic status, idioms, proverbs, graphical variants, abbreviations, names, derivations of names, onomatopoeic words. Typically, however, the lexical item is a lexeme, an abstract unit in the semantic system of a language which subsumes inflectional variation. The lemma thus subsumes the variant forms of the lexical item. These include variant spellings of the base form, as well as oblique forms in all their variant spellings. Lemmatisation is, thus, a process of classification.

4.1 Lexical Variation

Since lemmatisation concerns the classification of the variant forms of a

lexical item under their lemma, it is important to understand the nature of lexical variation. Within the Corpus of Cornish, two types of lexical variation can be found. Synchronic variation is concerned with the forms by which a lexeme is realised at one particular point in history. Diachronic variation, on the other hand, is concerned with variations in the forms by which a lexeme is realised over a period of time. Diachronic variation is thus allied to etymology.

4.1.1 Synchronic variation

In Cornish, synchronic lexical variation may be inflectional, derivational, conditioned or free (see Figure 25). In Cornish, countable nouns, verbs, prepositions, adjectives and cardinal numbers may be inflected. Inflectional variation is realised paradigmatically by the concatenation of affixes with the stem. The lexicographer should take care to indicate any irregularities in the paradigm and make explicit in the lemma to what declension or conjugation an item belongs. Derivational variation is also realised by affixation but may result in a change of word class and consequently a new inflectional paradigm. The boundary between derivation and inflection is not always transparent. The lexicographer must decide whether to give derivatives full entries, include them in the entry for the main form, or omit them entirely from the word list. Conditioned variation includes mutation of initial consonants and apocope. Free variation is the result of the non-systemic substitution of graphemes within the item which result in no change of meaning. Free variation of the base form entails problems regarding whether variant spellings of the base form should be given separate entries or cross referenced to their canonical

form.

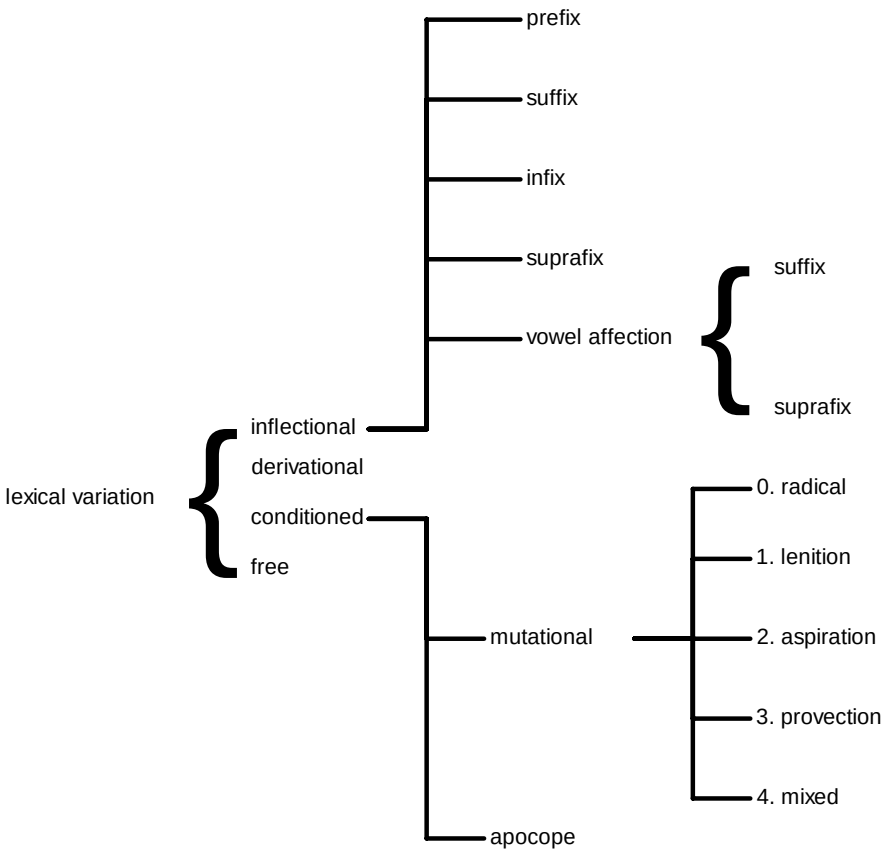


Figure 25 The synchronic variation system of Cornish

The lemma conventionally represents all the inflected forms of the lexical item which are then normally treated together in the same entry, and under the same entry form. An oblique form is one of any of the forms of a lexeme except its base form. In the corpus of Cornish and the dictionaries derived there from, the oblique forms of the lexeme include not only inflected forms but also conditioned variants and free variants.

For some languages, some dictionaries give variant forms, especially the new

dictionaries with their information for encoding. They may give only the irregular variants, or they may give all. Alternatively variants may be omitted from the dictionary, in which case the user is expected to retrieve them with the help of a grammar (Béjoint 1994: 192). No ambiguity should exist concerning the paradigm of the whole lexeme and Zgusta (1971: 121) recommends that any irregularities be indicated in the lemma. In addition, irregular inflected forms should normally be given a separate entry, with usually only a cross-reference to the main entry for the lemma (Béjoint 1994: 192). Sometimes the base form could belong to different paradigms. In this situation, Zgusta (1971: 121) recommends that the lexicographer indicates, "such information as makes the rest of the paradigm clear and unambiguous. ... if a good number of the canonical forms of a language requires further specifications and indications in order to yield the paradigm unambiguously, the lexicographer will do well to supply these indications everywhere". In the case of lexemes whose paradigms are not apparent from their base form, reference to the appropriate paradigm is necessary. This can be achieved by a numbering system or by cardinal forms and indications which unambiguously implicate the paradigm (Zgusta 1971: 121).

Zgusta (1971: 122) points out that in addition to irregular forms, the irregular absence of forms also needs to be indicated. If it is the base form that is missing, the lexicographer will have to select another form as the canonical form.

Inflection involves a change in the form of the word that signals a change in the grammatical category but leaves the word's lexical meaning

unchanged. Only grammatical inflection may be strictly regarded as formal variation of the word. The lexicographer considers all members of a single paradigm to be variant forms of a lexeme and, therefore, uses the canonical form to represent the single entry for that lexeme (Zgusta 1971: 127-31).

The basic precept that a relationship exists between words such as *lyver* (‘book’) and *lyfrow* (‘books’) or *obery* (‘work’) and *oberys* (‘worked’) lies at the heart of lexicography. This word-and-paradigm model entails the lexical meaning of the items comprising the paradigm being identical. The different forms thus represent only grammatical differences (Zgusta 1971: 119). The semantic unit which underlies the paradigm is referred to as the lexeme. In Cornish, countable nouns, verbs, pronominal prepositions, adjectives and cardinal numbers may be inflected. Figure 26 shows a system network of Cornish inflection.

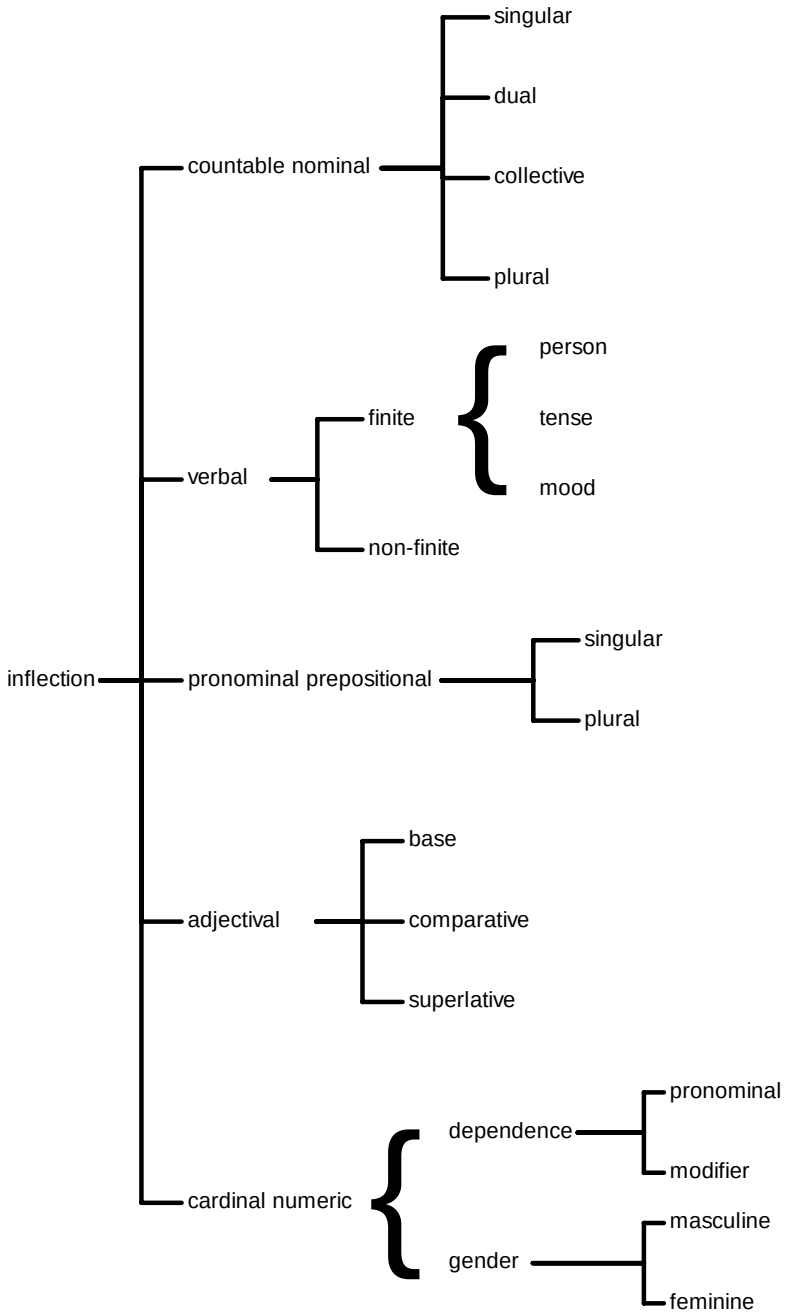


Figure 26 The Cornish inflection system

One problem that the lexicographer may encounter is in deciding if s/he is dealing with a single paradigm. One test is to see if the entire paradigm of an item really has the same lexical meaning. Sometimes polysemous words are

encountered the different senses of which may be partially distinguished by differences in their paradigm. Sometimes items are encountered where one form of the paradigm displays a peculiarity in its lexical meaning that is not present in the lexical meaning of the other forms. Lexicographic and grammatical tradition are likely to influence the lexicographer in his decision whether or not to treat a set as a single lexeme (Zgusta 1971: 123-7).

Suppletion is the grammar's use of a form with a different root to complete a paradigm. A rule based relationship between morphemes is, thus, not apparent. Zgusta (1971: 123) recommends that suppletives be, not only, listed in the lemma, but also given separate entries with a cross reference to the canonical form. Thus we find *gwell* given as the comparative of the adjective DA ('good') (NCED, CED). Similarly *tus* is commonly given as the plural of DEN ('man') by 20th century Cornish lexicographers (NCED, CED; GKK, NSCD). Nevertheless DEN has the regular plural attestations "dens", "denes" (ACB) and "dynion" (AB) in Modern Cornish. We similarly find the suppletive comparatives *gweth* and *lacca* as well as the regular comparative *drocca* given for the adjective DROG ('bad') (NCED, CED). In the case of the adjective DA, the suppletive form *gwell* is needed to complete the paradigm. In the cases of *tus*, *gweth* and *lacca*, they appear to have been treated as suppletive forms of DEN and DROG because they lack attested base forms.

In Cornish, countable nouns may be singular, collective, dual, singulative or plural. Figure 27 shows a system network of nominal inflection in Cornish. Singular and collective nouns are in the uninflected base form. Dual,

singulative and plural nouns are all inflected.

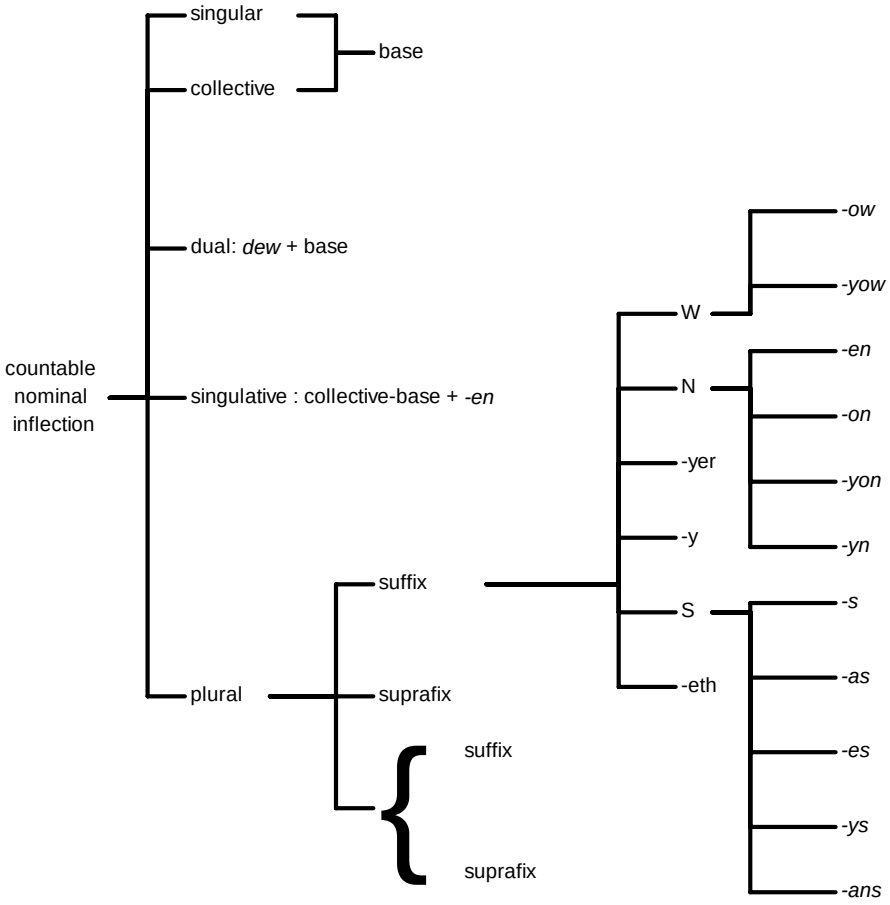


Figure 27 The Cornish nominal inflection system

The dual noun is formed by adding the prefix *dew-*, *diw-*, *deu-*, *du-*, *dyu-* or *dyw-* to the base form of a noun. Thus the noun, "lagas" (*Origo Mundi*: line 1109), 'an eye', has the dual form, "dewlagas" (*Passio Domini*: line 396), 'a pair of eyes'. Similarly the noun, "lef" (*Origo Mundi*: line 587), 'a hand', has the dual form, "dyulef" (*Passio Domini*: line 2375), 'a pair of hands'.

Collective nouns denote a group or class and usually refer to natural objects

such as plants or animals. Examples of collective nouns include “gwedh” (*Pascon Agan Arluth*: stanza 16), ‘trees’, and “nêdh” (AB: 78), ‘nits’. The singulative is formed from a collective noun by the addition of the suffix, *-an*, *-enn* or *-en*. Thus the collective noun, “gwedh”, ‘trees’, has the singulative form “gwedhenn” (*Origo Mundi*: line 29), ‘a tree’. Similarly the collective noun, “nêdh”, ‘nits’, has the singulative form “nedhan” (AB: 78), ‘a nit’.

The plural is formed by the addition of either a suffix, an infix or a suprafix to the base. In addition, suffixes may be combined with a vowel affection suprafix. Examination of the corpus reveals six basic types of plural noun suffix in Cornish, the W type, the N type, -YER, -Y, the S type and -ETH. Of these, the W, N and S types can be further subdivided. Nouns which take the W type of plural suffix may end in either -YOW or -OW. Thus “tyr” (*Passio Domini*: line 392), ‘land’ has the plural form, “tyryow” (*Origo Mundi*: line 26), ‘lands’. Similarly “fos” (*Origo Mundi*: line 1690), ‘a wall’, has the plural form “fosow” (*Origo Mundi*: line 2320), ‘walls’. If the singular ends in a consonant, this is sometimes found doubled. Thus we also find the form “fossow” (*Origo Mundi*: line 2450), ‘walls’. The N type can be subdivided into the suffixes -EN, -ON, -YON and YN. N type suffixes are frequently accompanied by vowel affection suprafixation of the final vowel of the base form. Thus “ky” (*Resurrexio Domini*: line 2026), ‘a dog’, has the plural form “kuen” (*Resurrexio Domini*: line 172), ‘dogs’; “lader” (*Pascon Agan Arluth*: stanza 186), ‘a thief’, has the plural form, “laddron” (*Pascon Agan Arluth*: stanza 186), ‘thieves’; “mab” (*Pascon Agan Arluth*: stanza 1), ‘a son’, has the plural form “mebyon” (*Origo Mundi*: line 1038), ‘sons’; “box” (VC), ‘a box

tree', has the plural, "byxyn", 'box trees' (VC). The noun "pren" (*Origo Mundi*: line 1444), 'a piece of timber' takes the *-yer* suffix to form the plural "prennyer" (*Pascon Agan Arluth*: stanza 151). The noun "mowes" (*Passio Domini*: line 1876), 'a girl', takes the suffix *-y* to form the plural "mowysy" (*Passio Domini*: line 944). The S type can be subdivided into the suffixes -S, -AS, -ES, -YS and -ANS. Thus "person" (*Origo Mundi*: line 1771), 'a person', has the plural "persons" (*Origo Mundi*: line 110), 'people'; "floch" (*Origo Mundi*: line 390), 'a child' has the plural forms "flehas" (*Passio Domini*: line 1168) and "fleghys" (*Origo Mundi*: line 1585), 'children'; "best" (*Origo Mundi*: line 124), 'an animal' has the plural form, "bestes" (*Origo Mundi*: line 42), 'animals'; "car" (*Pascon Agan Arluth*: stanza 93), 'a friend', has the plural form "kerans" (AB: 50), 'friends'. The noun "el" (*Resurrexio Domini*: line 787), 'an angel', takes the suffix -ETH to form the plural, "eleth" (*Resurrexio Domini*: line 190), 'angels'.

Some nouns form their plural by vowel affection only. Thus "tros" (*Passio Domini*: line 2781), 'a foot', has the plural form "treys" (*Passio Domini*: line 2937) 'feet'; "men" (*Passio Domini*: line 3211), 'a stone', has the plural form "meyn" (*Passio Domini*: line 62), 'stones'; "broder" (*Origo Mundi*: line 127), 'a brother', has the plural form "breder" (*Resurrexio Domini*: line 1163), 'brothers'; "daves" (*Origo Mundi*: line 127), 'a sheep', has the plural form "deves" (*Origo Mundi*: line 1065), 'sheep'; "ascorn" (*Resurrexio Domini*: line 2598), 'a bone' has the plural forms "escarn" (*Origo Mundi*: line 2743) and "yscarn" (*Passio Domini*: line 3173), 'bones'. These suprafixes are shown by the system network in Figure 28.

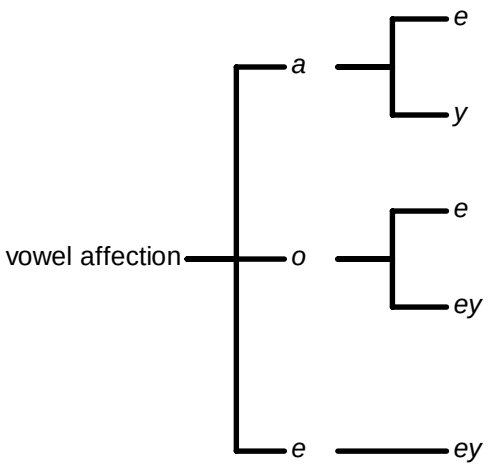


Figure 28 The vowel affection system

Cornish verbs may be either finite or non-finite. Finite verbs are inflected for person, tense and mood. Inflection for person includes first, second and third persons singular, first, second and third persons plural, and an impersonal form. There are four tenses: present, imperfect, preterite and pluperfect. In addition, Cornish has three moods: indicative, subjunctive and imperative. Figure 29 shows a system network for verbal inflection in Cornish. The stem of the verb is the third person singular of the present indicative or the second person singular of the imperative, these two having one and the same form.

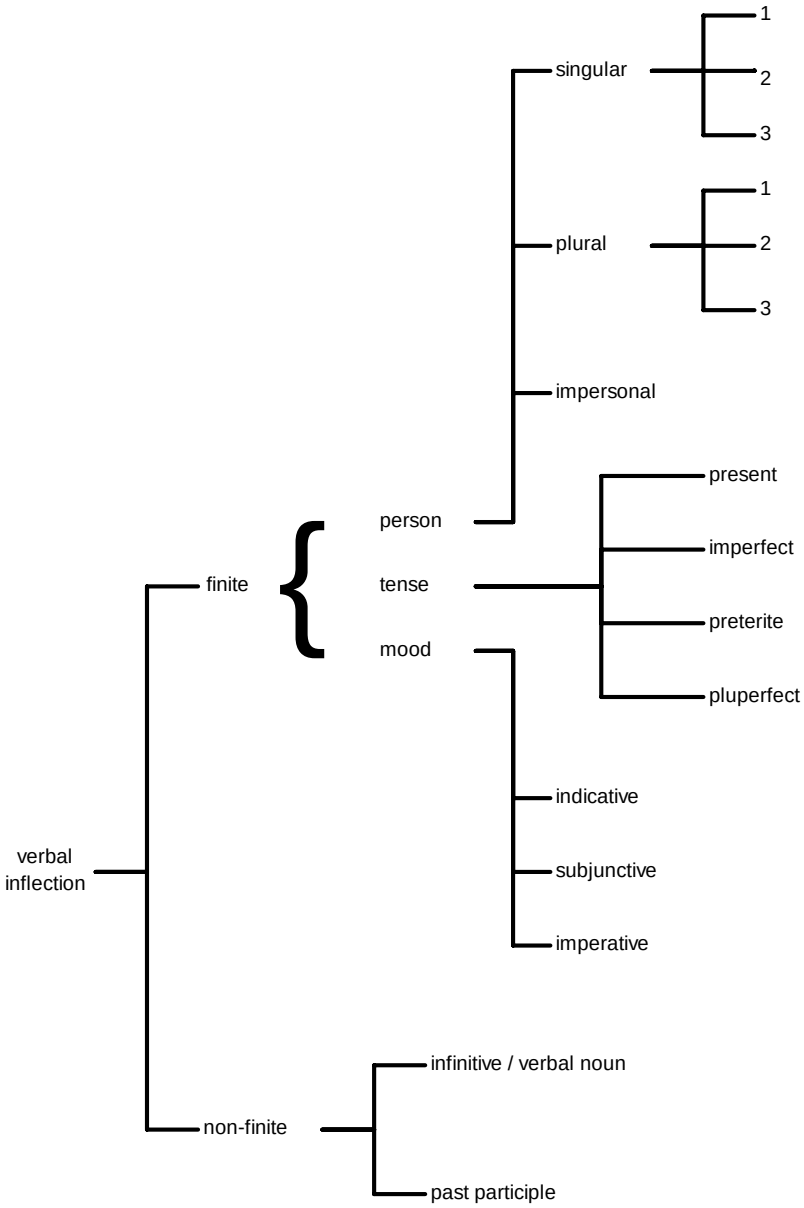


Figure 29 The verbal inflection system

The Cornish infinitive or verbal noun is formed by adding a suffix to the stem of the verb. In Middle Cornish, five regular infinitive suffixes are found: -E, -Y, -S, -L, and -N. Thus we find “care” (*Origo Mundi*: line 1126), ‘to love’; “dybry” (*Origo Mundi*: line 264), ‘to eat’; “myras” (*Origo Mundi*: line 1412), ‘to see’; “leverel” (*Passio Domini*: line 1759), ‘to say’; “danfon” (*Passio*

Domini: line 1615), ‘to send’. Lhuyd (AB: 245), writing in the Modern Cornish Period, also notes five regular infinitive suffixes in Cornish: -A, -I or -Y, -S or -Z, -L, and -N. There are also a small number of Cornish verbs, such as CUNTELL, for which the stem is identical with the infinitive.

The Cornish past participle is formed by adding a suffix to the stem of the verb. In Middle Cornish, -YS is by far the most common past participle suffix. Thus “gorrys” (*Resurrexio Domini*: line 430) is the past participle of GORRA, ‘put’. A vowel affection suprafix may also co-occur with the suffix. Thus “kyrys” (*Resurrexio Domini*: line 892), ‘loved’, is the past participle of CARA, ‘to love’. However -IS and -ES are also attested. Thus “kefis” (*Pascon Agan Arluth*: stanza 151) is found as the past participle of “cafos” (*Pascon Agan Arluth*: stanza 164), and “res” (*Passio Domini*: line 2496), ‘given’ is found as the past participle of “ry” (*Origo Mundi*: line 103), ‘give’ . Lhuyd (AB: 248), writing in the Modern Cornish period, notes vowel affection and three past participle suffixes, -YZ, -EZ and -AZ. Gendall (SDMC: 91) notes five main past participle suffixes that are attested in Modern Cornish, -EZ, -AZ, -EZE, -EIZ and -EAZ. Figure 30 shows a system network of past participle formation.

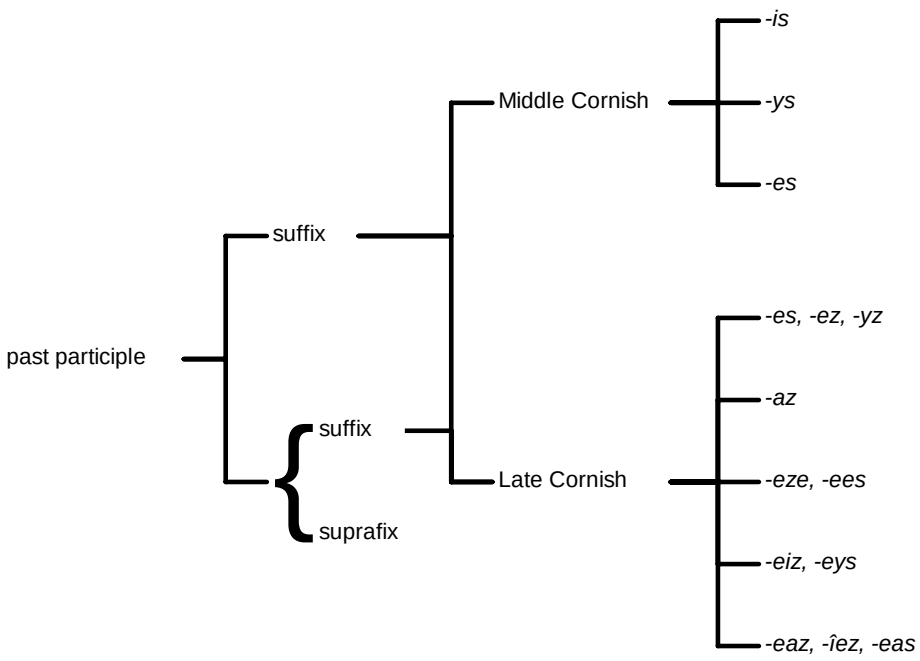


Figure 30 The past participle inflection system

Whilst in the case of English, paradigms may be fairly limited, in some languages the paradigm of a lexeme may include a considerable number of forms (Zgusta 1971: 119). Figure 31 shows the inflectional suffixes of regular verbs in Middle Cornish for person, tense and mood.

Indicative				
	Present	**Preterite**	**Imperfect**	**Pluperfect**
1 singular	*-af*	*-ys*	*-en*	*-sen*
2 singular	*-yth*	*-sys*	*-es*	*-ses*
3 singular	*-ø*	*-as*	*-e, -a*	*-se*
1 plural	*-yn*	*-syn*	*-en*	*-sen*
2 plural	*-ough, -eugh*	*-sough*	*-eugh*	*-seugh*
3 plural	*-ons*	*-sons, -sans*	*-ens*	*-sens*
0 impersonal	*-er, -yr*	*-as*	*-ys*	

Subjunctive		
	Pres./fut.	**Imperfect**
1 singular	*-yf*	*-en*
2 singular	*-y*	*-es*
3 singular	*-o*	*-e*
1 plural	*-yn*	*-en*
2 plural	*-eugh, -ough*	*-eugh*
3 plural	*-ons*	*-ens*
0 impersonal	*-er, -ser*	

Imperative
- ø
-es, -ens
-en, yn
-eugh, -ough
-ens, -es

Figure 31 The inflectional suffixes of regular verbs in Middle Cornish

The pronominal complement of a preposition is represented by a suffix attached to the base of the preposition. Figure 32 shows a system network of pronominal preposition suffixation.

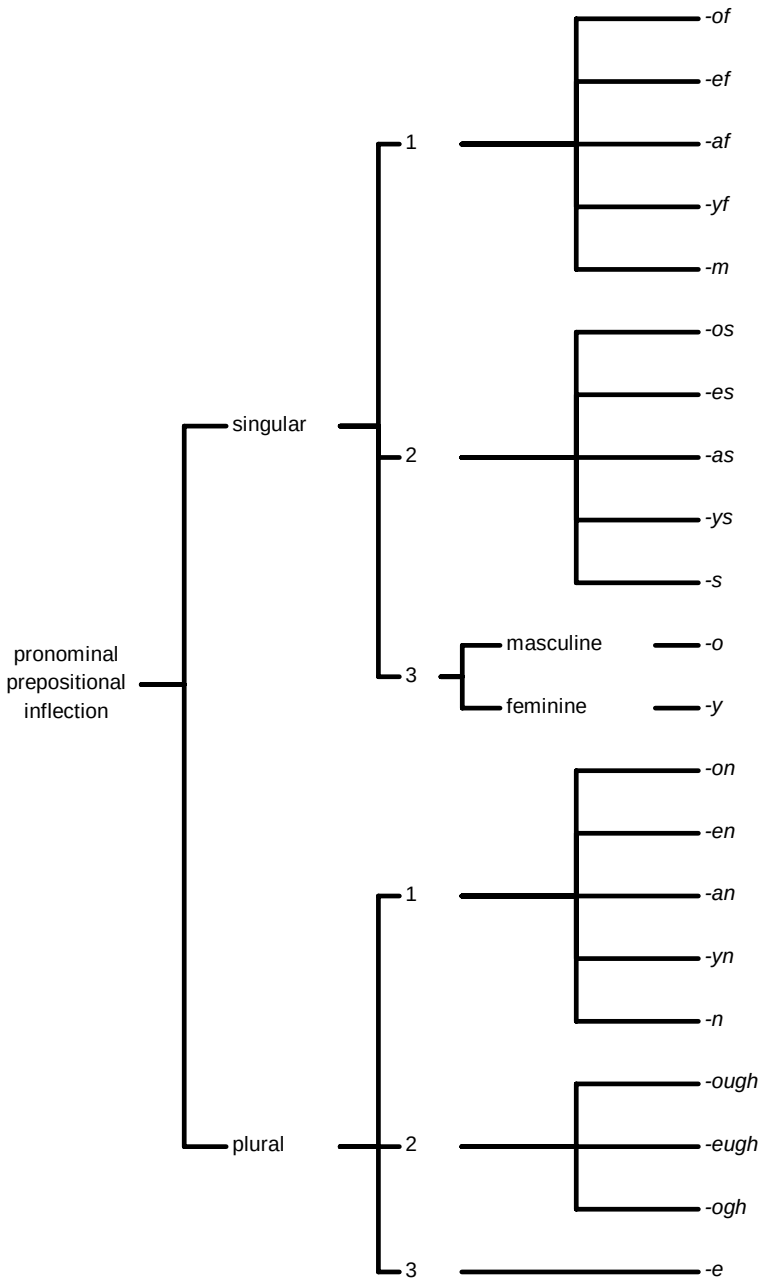


Figure 32 The pronominal prepositional inflection system

These pronominal suffixes frequently accompany an infix and/or a suprafix. A consonant may be doubled or changed, or a syllable may be added, or there may be vowel affection. Thus the preposition YN, ‘in’, has the pronominal form “ynnof” (*Resurrexio Domini*: line 707), ‘in me’; WORTH, ‘against’, has the pronominal form “worte” (*Origo Mundi*: line 2476) ‘against them’; WAR, ‘upon’, has the pronominal form “warnotho” (*Origo Mundi*: line 1539), ‘upon him’; the preposition DRE, ‘by’, has the pronominal form “drythy” (*Origo Mundi*: line 1668), ‘by her’. Some pronominal prepositions are sometimes found with an ending after the suffix. Thus THE, ‘to’, has the pronominal forms “thym” (*Origo Mundi*: line 2286) and “thymmo” (*Origo Mundi*: line 2256), ‘to me’.

The comparative and superlative forms of the adjective are marked by inflection and are usually both formed by the addition of the suffix -A, -HA, -E or -HE. Thus “pell” (*Pascon Agan Arluth*: stanza 160), ‘far’, has the comparative form “pelha” (*Gwavas Manuscripts*: 103r), ‘further’. In addition to the suffix, there may be a doubling of the final consonant of the base. Thus “tek” (*Pascon Agan Arluth*: stanza 161), ‘fair’, has the comparative form “tekke” (*Pascon Agan Arluth*: stanza 226), ‘fairer’; and “uhel” (*Origo Mundi*: line 805) has the superlative form “uhella” (*Passio Domini*: line 2189). There may be vowel affection or vowel elision and the final consonant of the base may be devoiced. Thus “hager” (*Pascon Agan Arluth*: stanza 122), ‘hideous’, has the comparative form “haccra” (*Pascon Agan Arluth*: stanza 151), ‘more hideous’.

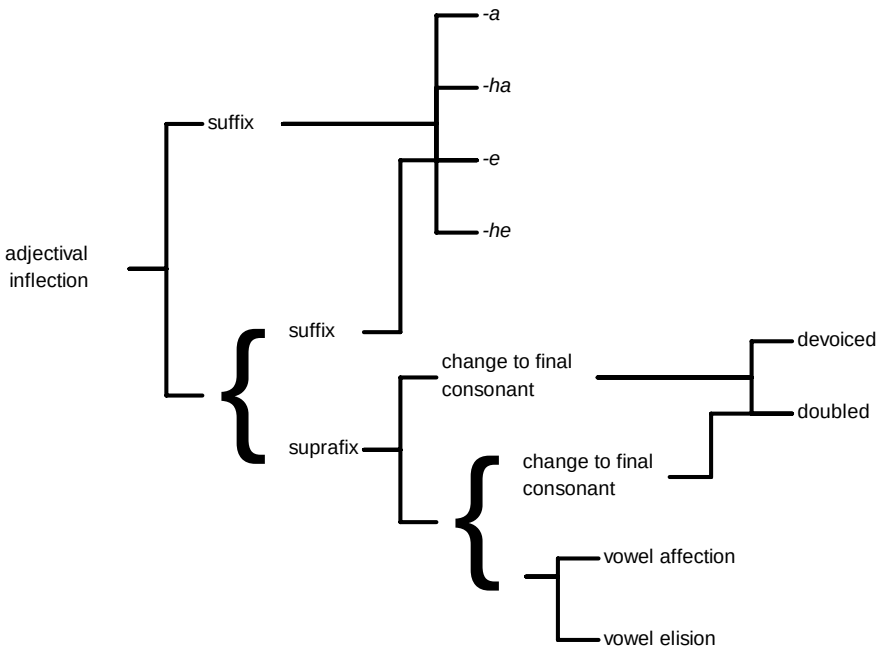


Figure 33 shows a system network of adjectival inflection in Cornish.

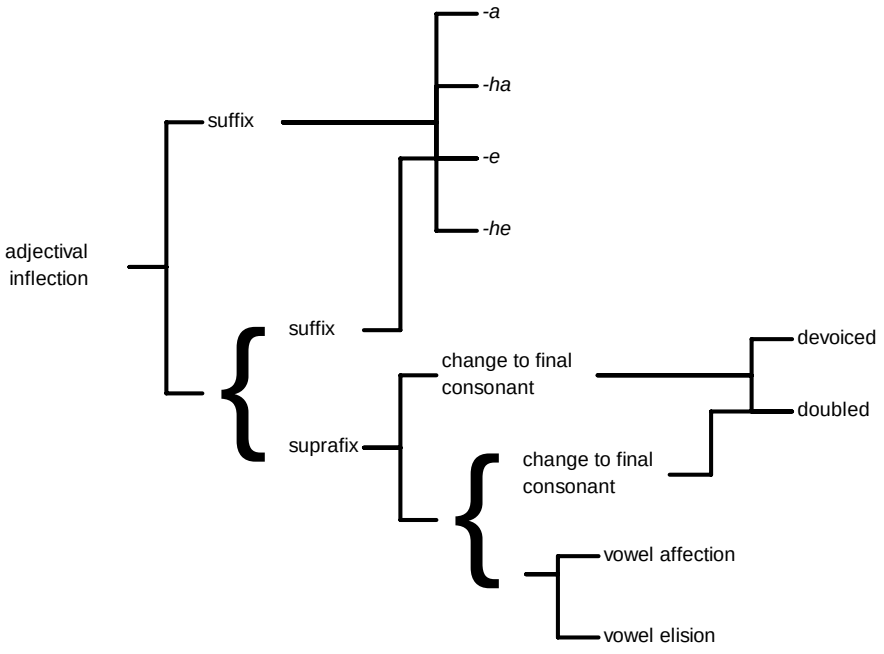


Figure 33 System network of adjectival inflection in Cornish

There is a tradition of suppletion to complete certain adjectival paradigms in Cornish. According to Lhuyd (AB: 243), *ogoz*, 'near', has the comparative "nêz", 'nearer', and the superlative "nèsa", 'the next'. Williams (LCB: 265-6) writes that *nes*, 'nearer, near, again', is used as a comparative of *agos*, 'near', and that *nessa*, 'nearest, next, hithermost, second' is used for the superlative of *agos*. Jenner (1904: 93) writes that *nes* and *nessa* are comparative and superlative of *ogas*. Lewis (1990) gives *nes* and *nessa* as comparative and superlative of *agos*.

The comparative and superlative forms, *gwell* and *gwella* have a choice of two possible suppletive base forms. According to Lhuyd (AB: 243), *gụel*, 'better', and *gụèla*, 'best', are comparative and superlative of *mâz*, 'good'. Williams (LCB: 195) writes that *gwell*, 'better', is the irregular comparative of *da*, 'good', or *mâs*, 'good', and that *gwella* is the irregular superlative of *da*, or *mâs*. Jenner (1904: 93) writes that *gwel* and *gwella* are the comparative and superlative forms of *da* or *mas*. Lewis (1990: 18) gives *guel* and *guella* as the comparative and superlative of *da* or *mas*.

Drok is shared as a suppletive base form for two different comparative-superlative pairs and also has its own superlative, *droca*. According to Lhuyd (AB: 243), "drok", 'bad', has the comparative "gụaeth", 'worse', and the superlatives "dròka" or "gụaetha", 'worst'. Williams (LCB: 198) writes that *gweth*, 'worse', is used as the comparative of *drôc* and that *gwetha*, 'worst' is used as the superlative of *drôc*. However, Williams (LCB: 114) also gives *droca*, 'worst'. Jenner (1904: 93) writes that *gwêth* and *gwêtha* are

comparative and superlative forms of *drôg*, but usually *lakkah*, comparative of *lak*, is used to signify ‘worse’. Lewis (1990: 18) gives *gueth* and *guetha* as the comparativeand superlative of *drok*.

Bîan, *bechan* or *behan* is used as a suppletive base form for the comparative, *le* and superlative, *leiha*, but has its own comparative and superlative forms. According to Lhuyd (AB: 243), “bîan”, ‘small’, has the comparative “le”, ‘less’, and the superlative “leiha”, ‘least’. Williams (LCB: 231) writes that *le*, ‘less, smaller’, is used as the comparative of *bechan*, and that *leia*, ‘least’ is used as the superlative of *bechan*. Williams also has an entry for *behan*, ‘little, small’, which he says is another form of *bechan* and has the comparative form *behannah*. Jenner (1904: 93) writes that *leh* and *lŷha* are comparative and superlative forms of *bîan*, but that there is also a comparative, *behadnah*, and superlative, *behadna*. Lewis (1990: 18) gives *le* and *lyha* as the comparative and superlative of *beghan*.

Some cardinal numbers are inflected for gender and one number is inflected for dependence. Figure 34 shows a system network of cardinal numeric inflection. The number 3 has the masculine form “try” (*Resurrexio Domini*: line 374) and the feminine forms “tyr” (*Origo Mundi*: line 828), “tyyr” (*Origo Mundi*: line 1729) and “ter” (*Passio Domini*: line 147). Similarly the number 4 has the masculine form “peswar” (*Resurrexio Domini*: line 563) and the feminine form “pedyr” (*Origo Mundi*: line 772).

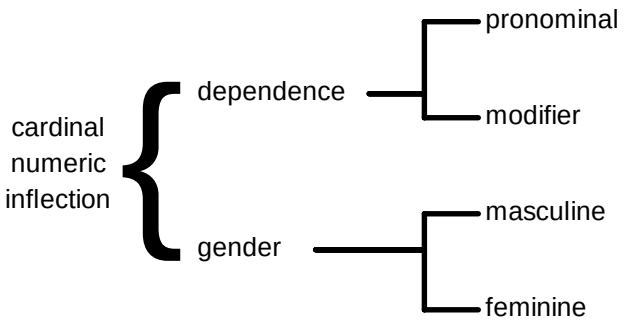


Figure 34 The cardinal numeric inflection system

Some writers and lexicographers (LCB 99, 117; 128; ECD2: 119; NCED: 39, 50; Smith 1972: 62; SDMC: 28; Brown 1993: 68; GKK: 71, 79) maintain that the number 2 has both masculine and feminine forms. However this distinction is not born out by attestation. In *Pascon Agan Arluth* only one form, *dew*, is attested for number 2. In the *Ordinalia* two forms are attested, *dew* and *dyw*. However they are not distinguished by gender. Thus we find the feminine noun *luef*, 'a hand' collocating with both forms, "dyw luef" (*Origo Mundi*: line 1346) and "dew luef" (*Origo Mundi*: line 1534); we find the masculine noun *dorn*, 'a fist', collocating with "dyw" (*Resurrexio Domini*: line 2178) and the masculine noun *adla*, 'a rogue', collocating with "dew" (*Resurrexio Domini*: line 1479). Jordan (*Gwreans an Bys*) uses three forms *deaw*, *dew* and *thyw*. All three are used for both masculine and feminine. Thus we find both the feminine noun "gweth" (*Gwreans an Bys*: line 966), 'a garment', and the masculine noun "vabe" (*Gwreans an Bys*: lines 1054, 1232), 'a son' collocating with *deaw*; we find both the feminine noun "wreag" (*Gwreans an Bys*: line 1452), 'a wife', and the masculine noun "ran" (*Gwreans an Bys*: line

1707), ‘a part’, collocating with *dew*; we find the masculine noun *fridg* or *freyge*, ‘nostril’, collocating both with “thyw” (*Gwreans an Bys*: line 1854) and with “thew” (*Gwreans an Bys*: line 1933) .

The number 1 is inflected for dependence. When ‘1’ is a modifier in a nominal group it has the form “un” (*Passio Domini*: line 160). ‘1’ also has the pronominal form “unan” (*Charter Endorsement*: line 7), “onan” (*Origo Mundi*: line 3), “onon” (*Resurrexio Domini*: line 1403) or “onyn” (*Gwreans an Bys*: line 142). When *onan* is pre-modified by a possessive pronoun it has the form “honan” (*Pascon Agan Arluth*: stanza 6), “honon” (*Pascon Agan Arluth*: stanza 101) or “honyn” (*Gwreans an Bys*: line 1527).

Variation of the word form that is partly or wholly determined by linguistic context is said to be conditioned (Crystal 1985: 64). There are two types of conditioned variation found in Cornish, mutational variation and apocope.

The synchronic mutational system affects the initial consonant of a word in certain grammatical situations. These changes are referred to as mutations. By convention, mutations are normally classified under four or five main states: the radical or unmutated state, lenition, aspiration, provection and mixed (AB: 241-3; Norris 1859b 9-12; Jenner 1904: 68-72; Allin-Collins 1927: 7-8; Smith 1972: 14; George 1986: 77; Lewis 1990: 7-10; SDMC: 140-1; Brown 1993: 10). Figure 35 shows a system network of synchronic mutational variation.

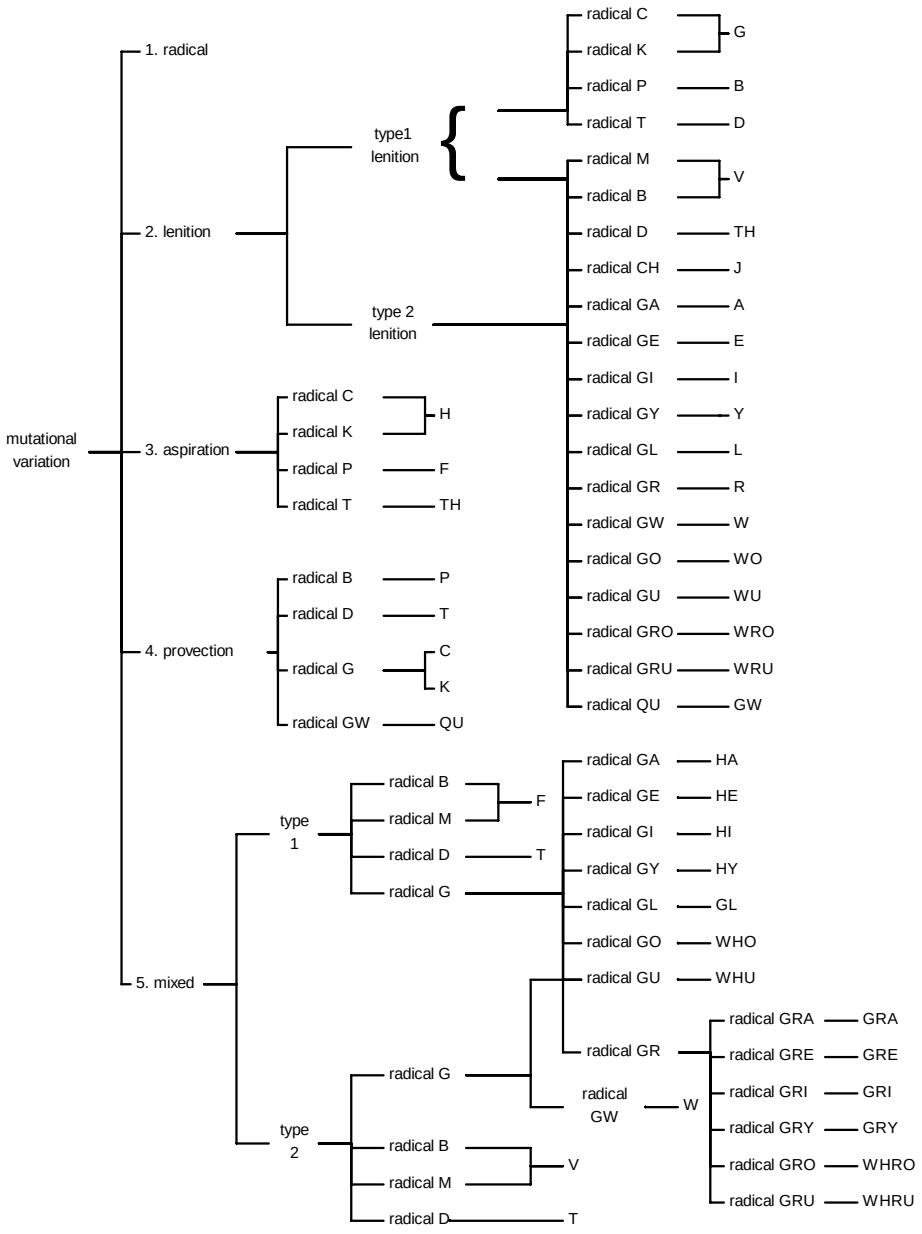


Figure 35 The synchronic mutational variation system

In the radical state, the initial letter is that of the canonical form. Lemmatisation thus involves changing the initial consonant of the other four mutational states to the radical state.

Lenition, the second state mutation, can be subdivided into two types, of which, type 2 is a subset of type one (see Figure 35). In addition to the lenitions shown in Figure 35, in Modern Cornish, the initial consonant <F> lenites to <V> and the initial consonant <S> lenites to <Z>. Thus Wella Kerew (*Gwavas Manuscripts*: 104r) writes “teeze veer”, ‘wise men’, and “an zettyas”, ‘set him’. These Modern Cornish mutations are sometimes referred to as “new lenition“ (Jackson 1967: 497-519; George 1986: 78-9).

Figure 36 shows a system network of the conditions under which lenition is found. These conditions include the word preceding the mutated word, the word following the mutated word and the mutated word itself. There is some variation to be found between texts however. Smith (1984: 38) observes that *a pup*, ‘of each’, *dhe pup*, ‘to each’, and *war pup*, ‘on each’, are never mutated in the *Ordinalia*, but are always mutated in *Pascon Agan Arluth*.

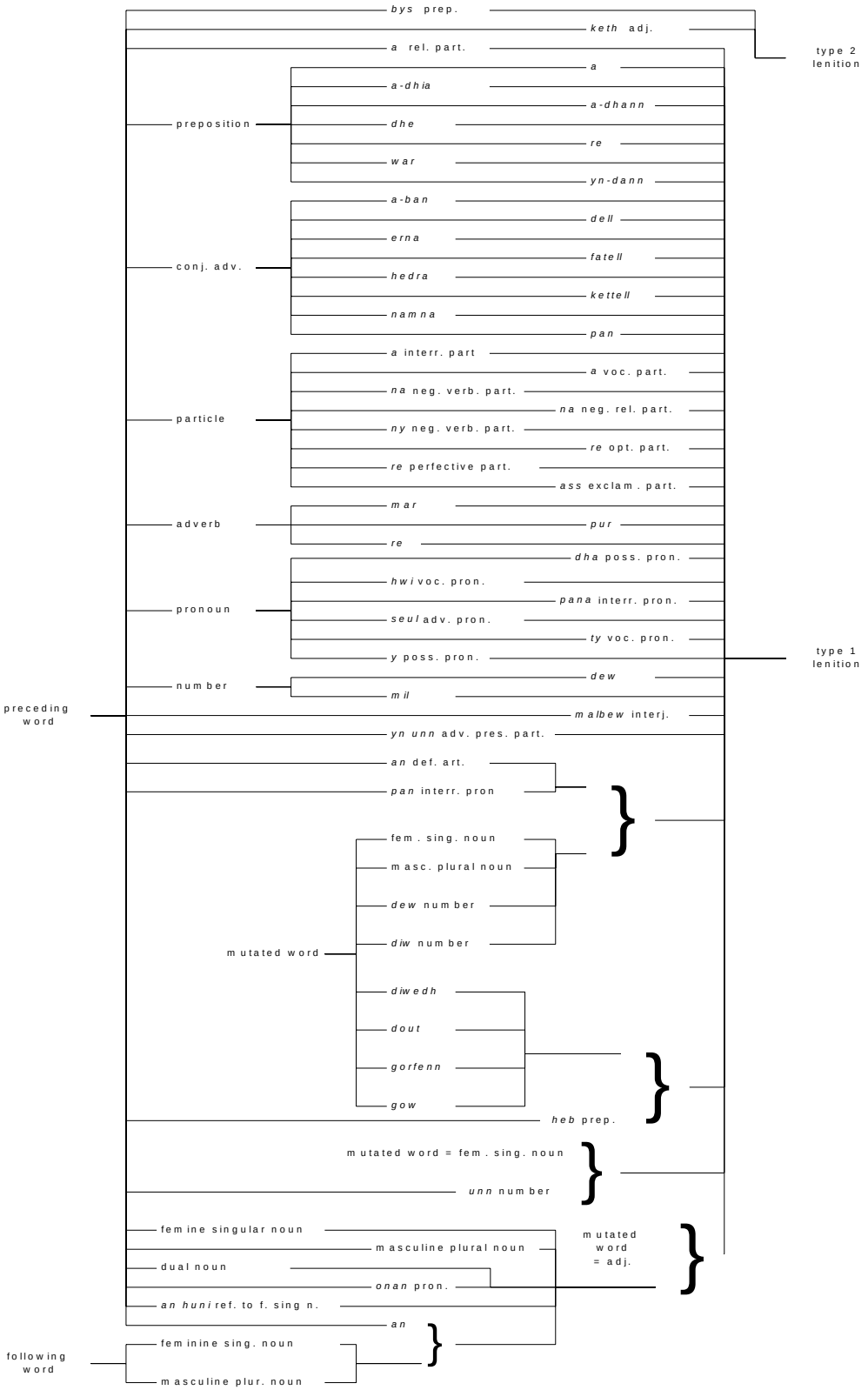


Figure 36 The causes of lenition system

The new lenitions found in Modern Cornish occur under similar but slightly different conditions from those shown in

Figure 36. One of these differences is that new lenition effects adjectives following both genders of noun. Another difference is that lenition of <F> and <S> after the definite article, *an*, is not restricted to feminine nouns.

Figure 37 shows a system network of the conditions under which aspiration, the third state mutation, is found.

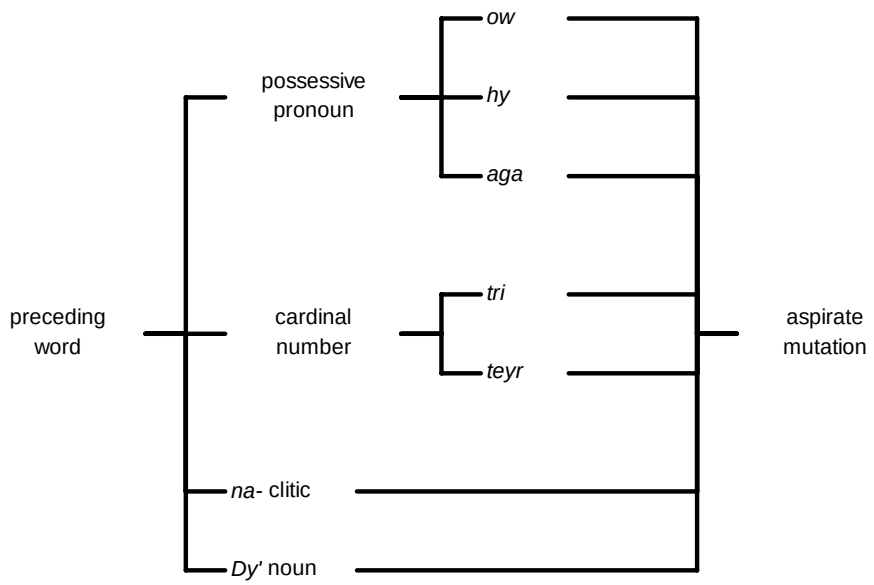


Figure 37 The causes of aspiration system

Figure 38 shows a system network of the conditions under which provection, the fourth state mutation, is found.

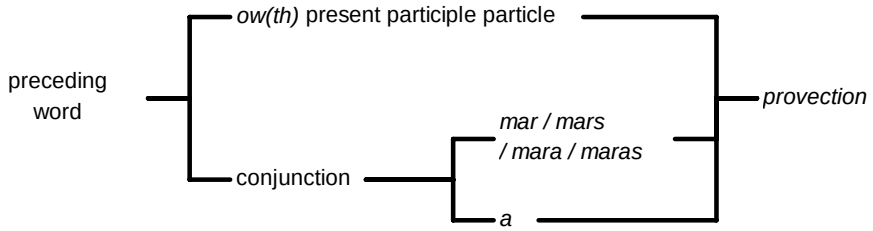


Figure 38 The causes of provection system

Figure 39 shows a system network of the conditions under which mixed, the fifth state mutation, is found.

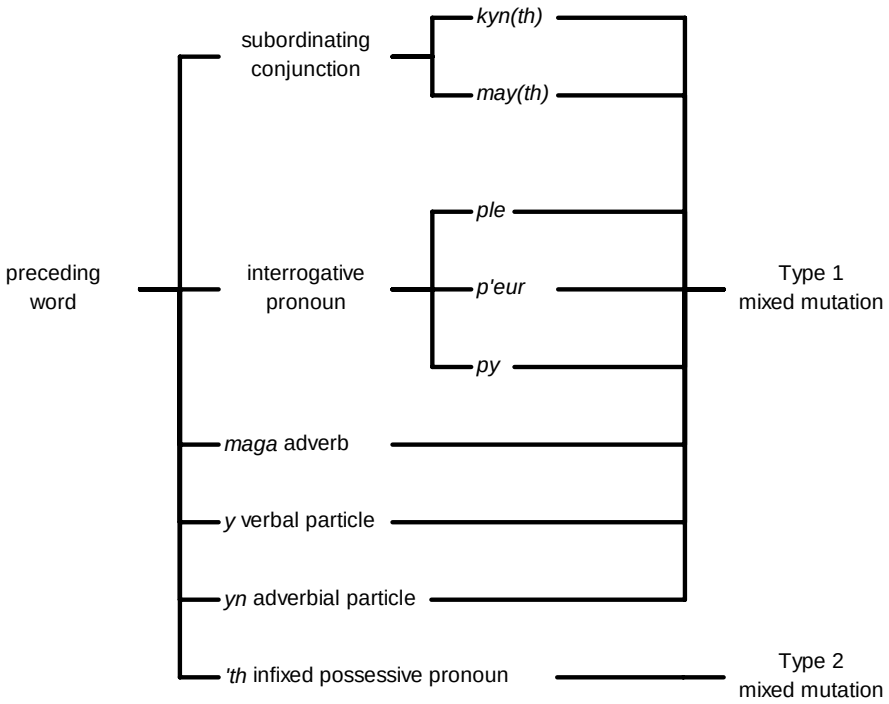


Figure 39 The causes of mixed mutation system

Mutations are not always strictly observed in the texts. Smith (1984: 38) notes

that mutations are more rigidly observed in *Pascon Agan Arluth* and *Gwreans an Bys* than in the *Ordinalia* and *Beunans Meriasek*. Figure 40 shows frequencies of missed mutations in these texts.

Pascon Agan Arluth		1 in every 74 lines
Ordinalia	*Origo Mundi*	1 in every 21.5 lines
	Passio Domini	1 in every 11.333 lines
	Resurrexio Domini	1 in every 10.666 lines
Beunans Meriasek		1 in every 9.666 lines
Gwreans an Bys		1 in every 48 lines

Figure 40 Frequencies of missed mutations in the corpus

Certain words have alternate terminations in which the final consonant or consonant cluster is omitted if the following word begins with a consonant. This phenomenon is known as apocope. Thus we find “resons mar fol ha mar dyn” (*Pascon Agan Arluth*: stanza 100), ‘such foolish and cruel reasons’, and “an scornye hag an gweska” (*Pascon Agan Arluth*: stanza 114), ‘mocked him and beat him’. Figure 41 shows a system network of the words and terminations which are affected by apocope.

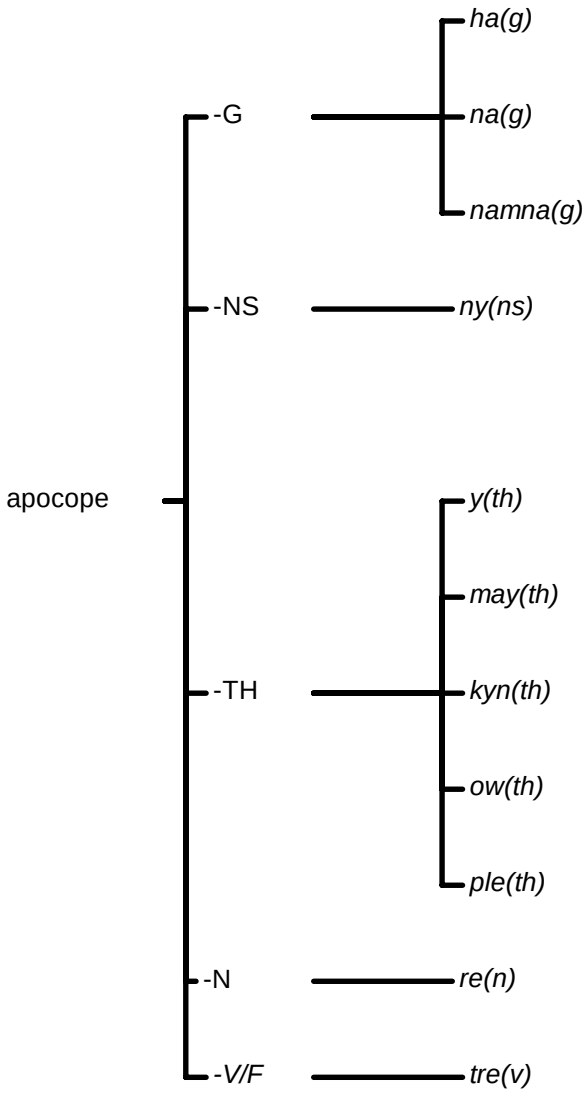


Figure 41 The apocope system

Sometimes words have parallel, different forms without any apparent difference in meaning. These, too, need to be indicated if the dictionary is descriptive (Zgusta 1971: 122). This is particularly prevalent in the corpus of Cornish in which there is a great deal of free variation of spelling. A case in point is the Cornish noun ALWETH (‘key’), which has the plural attestations “alwethow” (*Resurrexio Domini*: line 84), “alwhethow” (*Resurrexio Domini*:

line 634) and “alwheow” (*Resurrexio Domini*: line 650).

Occasionally a grapheme <H> or <W> is found at the beginning of words which otherwise begin with a vowel. Thus in *Pascon Agan Arluth* we find both “han **e**zewon ol adro” (*Pascon Agan Arluth*: stanza 146), ‘and the Jews all around’ and “han **h**uthewon ny wozye” (*Pascon Agan Arluth*: stanza 152), ‘and the Jews did not know’. And in the *Ordinalia* we find synonymous phrases variously spelled “an avel **w**orth y derry” (*Origo Mundi*: line 279) and “an avel **o**rth y dyrry” (*Origo Mundi*: line 195), ‘by plucking the apple’. Presence or absence of initial <H> or <W> in these and other similar occurrences do not appear to be conditioned by linguistic context and may, therefore, be considered examples of free variation.

Osselton (1995: 83-92) describes how early English lexicographers were confounded by the variety of spellings of the base form. Cawdrey (TA) brackets variants together so that, for example, *ingine* and *engine* appear only under the letter I. Other 17th century English lexicographers continued the practice so that alphabetical order generally deviates by about 6-8%. Philips (NWEW) adopts the formula *x* or *y* for variant base forms, for example:

> ‘Indocility’, or ‘Indocibility’ (lat.) an
>
> unaptnesse to be taught or learn.

He sometimes provides cross-references where spellings are far apart, such as

> ‘‘A Hodge-poge’ or ‘Hotch-pot’ ... flesh cut to pieces, and sodden together with Herbs’(121-2).

Johnson (DEL) is often blamed for current conventions of British English spelling (Mencken 1923: 235; Wrenn 1949: 99; Sheard 1954: 309). However

rather than reflecting his own preferences, Johnson (DEL) appears to be recommending the spelling used by printers of his time (Sled & Kolb 1955: 33,137; Osselton 1995:83 ff.).

In the case of Cornish lexicography, Williams (LCB) gives full separate entries to the variant base forms of Cornish BOS - 'to be': **BONES**, **BOS** and **BOSA**. **BONES** and **BOSA** are cross-referenced to **BOS**. Williams gives no indication, however, of the variant forms *bones* and *bosa* under the entry for **BOS**.

4.1.2 Derivational variation

On the one hand, inflection involves a change in the form of the word that signals a change in the grammatical category but leaves the word's lexical meaning unchanged. Derivation, on the other hand, involves a change in the form of the word with usually a change in the lexical meaning of the word. Inflected forms and derivatives are usually distinguished; HAVALDER ('likeness') and HEVELEBY ('to liken') are, thus, derivatives of HAVAL ('like' *adj.*), not inflected forms. Occasionally, however, differences in derivation entail no difference in meaning. Thus HAVALDER, HEVELEP and HEVELEPTER may all be translated into English as 'likeness'.

A difficulty that may arise is distinguishing derivation from inflection in cases where derivation is very regular. Thus the boundary between derivation and inflection is not always clear. Tradition is likely to play a part in deciding between derivation and inflection (Zgusta 1971: 127-31). A case in point is the Cornish verbal noun. Cornish lexicographers traditionally use the

verbal noun as the canonical form for the verbal paradigm. However the verbal noun, like other nouns, is either masculine or feminine and is part of the nominal paradigm. Thus 20th century Cornish dictionaries (NCED, CED; GKK, NSCD) list SKYLA ('cause') as a verb, and also as a feminine noun, with the plural form *skylys*. Another example of the difficulty of distinguishing between derivation and inflection concerns the verbal adjective and the past participle which in Cornish share the same form. Thus 20th century Cornish dictionaries (NCED, CED; GKK, NSCD) list FLERYS ('stinking', 'fetid') as an adjective, though it is also the past participle of FLERYA ('to stink').

A derivative may be given full entry status, be included within the entry for the main form, or in the case of decoding dictionaries in particular, for a derivative whose meaning is transparent, be omitted from the dictionary altogether (see Figure 42). When derivatives are simply listed this normally occurs at the end of the entry. This distinguishes them from inflected forms which are typically listed at the beginning. Decisions may be based on a number of criteria: practical considerations of space, optimal usefulness for users, the nature of the lexical unit (Ilson 1984: 80; Geeraerts 1989: 289; Béjoint 1994: 192).

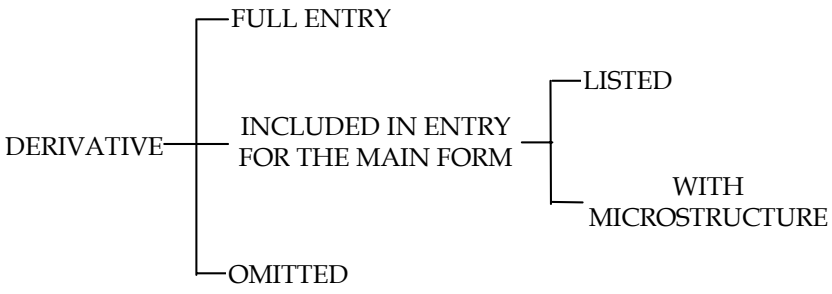


Figure 42 The derivative entry system

The lexicographer does not tend to indicate all derivations as separate entries, unless it is intended that a fully exhaustive dictionary be compiled. According to Zgusta (1971: 129),

> ... the greater the number of words in which the same derivational morpheme causes the same change in lexical meaning, the smaller will be the inclination of the lexicographer to list all these words. Put in another way, the more frequently a derivational morpheme can be used, and the more uniform its affect on the lexical meaning, the more does its function resemble a grammatical function. On the contrary, if the derivational morpheme is not frequent and/or if its modifying effect on the lexical meaning is far from uniform, the similarity to a grammatical function will be incomparably smaller and the lexicographer will be more inclined to indicate the respective words as separate items.

In Cornish, adverbs may be formed from adjectives by prefixing the particle YN-. Thus, from the adjective SUR (‘sure’), the adverb YN-SUR (‘assuredly’) is formed. This process is so regular in Cornish that it is unnecessary to list all of the adverbs thus formed in the dictionary.

The inclusion of derivatives in the entry for the main form may be used to indicate links between the base form and its derivatives. Such links are actually more concerned with form and morphological derivation than semantic relations (Béjoint 1994: 193). Indeed Hudson complains about the lack of correlation between lexical relatedness and the structure of traditional dictionaries. He points out that

> … to put two words in different lexical entries is to deny any connection between them, so that lexical relatedness is an all-or-none matter: either two words are related, in which case they will be shown in the same entry; or they are not related, and are in distinct entries.
>
> (Hudson 1988: 296-7)

It would seem logical to assume that all derivatives be included in the dictionary if it is intended as a complete record of the language. The lexicographer, however, has to decide whether 'all derivatives' refers to all and only those derivatives attested in the corpus, or whether to also include latent derivatives. The majority of dictionaries do not aspire to being so fully comprehensive since this would entail a vast number of entries. In practice, many derivatives which are semantically self-evident, do need to be included (Béjoint 1994: 192).

Landau challenges the assumption that certain forms are more basic semantically because they happen to be more basic grammatically.

> To regard adverbs ending in '-ly' or nouns in '-ness' as less important than the adjectival root indulges the convenience of the definer at the expense of the needs of the user. In many cases, '-ly' words and '-ness' words have acquired senses not adequately covered by the root words. '-ly' is supposed to mean 'in a -------- manner.' '-ness' is supposed to mean 'the state of being ----------'.... It is understood in theory, if not always observed in practice, that if a derivative has a meaning not covered by the senses of the form to which it is appended, or not applicable to the formulaic definitions cited, it should be entered separately and defined. As a result, adverbs like 'hopefully', 'incidentally', and 'literally' are defined as main entries in all reputable dictionaries.
>
> (Landau 1989: 78)

Thus the Cornish adverb YN-FAS ('well') is formed from the adjective MAS ('good') with mutation of initial <M> to <F>. However, when YN-FAS is used in the negative, it is translated into English as 'hardly', 'even' or 'scarcely'. Thus we find, "ny woffys yn fas un prygwyth genef golyas" (Passio Christi 1054) - 'couldn't you even keep watch with me for one moment'. Since this usage is unpredictable, it is necessary to list YN-FAS in

the dictionary.

A nonce is a form that a speaker invents or uses on a single occasion. The morphological process of derivation may employed for the creation of occasional nonces. Nonces and occasional forms are not usually included in a dictionary. However, it is very difficult for the lexicographer to distinguish occasional nonces and semi-stabilised forms from stabilised items (Zgusta 1971: 129-30). When one considers that nearly half of the word types attested in the corpus of Cornish are hapax legomena, one appreciates the difficulty in determining whether derivatives are stabilised.

4.1.3 Diachronic variation

Diachronic variation is concerned with variation of form found over a period of time. For example, the Old Cornish word “bochodoc” (VC), ‘poor’, is found as “bodjack” (*William Bodinar’s Letter*) in Modern Cornish. According to Trench (1857) a historical dictionary should record faithfully the older forms and spellings of words and give a full record of all the derivations. There are four types of diachronic variation: metathesis, intrusion, elision and mutation.

Metathesis is an alternation in the sequence of syllables or the transposition of phonemes (DLP2). Thus Middle Cornish “kepar” (*Pascon Agan Arluth*: stanza 123), ‘equally’, becomes “pekare” (*Gwreans an Bys*: line 2200) in Modern Cornish. Figure 43 shows some examples of metathesis between Middle and Modern Cornish.

Middle Cornish				Modern Cornish	
“ankevys”	(*Gwreans an Bys*: line 1346),	‘forgotten’	>	“neceaves”	(*William Bodinar’s Letter*)
“bolungeth”	*(Origo Mundi*: line 873),	‘the will’	>	“blonogath”	(*Gwreans an Bys*: line 95)
“dowr”	(*Pascon Agan Arluth*: stanza 58)	‘water	>	“dorrowe”	(*Gwreans an Bys*: line 2322)
“drehevell”	(*Pascon Agan Arluth*: stanza 203)	‘build’	>	“dereval”	(*Gwavas Manuscripts*:136v)
“drehevys”	(*Pascon Agan Arluth*: stanza 210)	‘built’	>	“dereves”	(*Gwavas Manuscripts*:103r)
“fatel”	(*Pascon Agan Arluth*: stanza 170)	‘how’	>	“fatla”	(*Gwreans an Bys*: line 2320)
“kenever”	(*Pascon Agan Arluth*: stanza 228)	‘each’	>	“kenevrah”	(*Gwavas Manuscripts*: 99v)
“yender”	(*Gwreans an Bys*: line 1667)	‘coldness’	>	“yeindre”	(*Gwavas Manuscripts*: 100r)

Figure 43 Metathesis between Middle and Modern Cornish

Intrusion refers to the addition of graphemes to an item. Intrusion can be of three types: prothesis, epenthesis and paragoge. Prothesis involves the insertion of a segment in word initial position. Prothesis is not very common in Cornish diachronic variation. An example is Middle Cornish “onour” (*Pascon Agan Arluth*: stanza: 136), ‘honour’, which becomes “honor” (*Gwreans an Bys*: line: 170) in Modern Cornish. Epenthesis, also called anaptyxis or svarabhakti, involves the insertion of a segment in word medial position. The insertion of a vowel into a cluster of consonants is frequently encountered. Thus Old Cornish “latro” (VC), ‘a thief’, becomes “lader” (*Pascon Agan Arluth*: stanza 186) in Middle Cornish. Figure 44 shows some examples of epenthesis between Middle and Modern Cornish. Paragoge refers

to the intrusion of a segment in word final position. Paragoge is not a common feature of Cornish diachronic variation. An example is Middle Cornish “huny” (*Pascon Agan Arluth*: stanza 240), ‘one’, which is found in Modern Cornish as “hunyth” (*Gwreans an Bys*: line 685) and “hunythe” (*Gwreans an Bys*: line 2248).

Middle Cornish				**Modern Cornish**	
“delyffrys”	(*Pascon Agan Arluth*: stanza 124)	‘release’	>	“delyverys”	(*Gwreans an Bys*: line 2464)
“dyffry”	(*Pascon Agan Arluth*: stanza 146)	‘indeed’	>	“devery”	(*Gwreans an Bys*: line 135)
“kyffrys”	(*Pascon Agan Arluth*: stanza 23)	‘likewise’	>	“keverys”	(*Gwreans an Bys*: line 955)
“lyffrow”	(*Pascon Agan Arluth*: stanza 17)	‘books’	>	“leverow”	(*Gwreans an Bys*: line 2176)
“ordna”	(*Pascon Agan Arluth*: stanza 7)	‘to order’	>	“ordayne”	(*Gwreans an Bys*: line 894)
“ple”	(*Pascon Agan Arluth*: stanza 147)	‘where	>	“peleah”	(*Gwavas Manuscripts*: 104r)

Figure 44 Epenthesis between Middle and Modern Cornish

Elision refers to the omission of a segment of an item. Both consonants and vowels may be affected and even entire syllables. There are three types of elision: aphesis, syncope and apocope. Aphesis, also known as prosiopesis, is the loss of an initial segment. Thus Middle Cornish “eseza” (*Pascon Agan Arluth*: stanza 13), ‘to sit’, becomes “zetha” (*Gwavas Manuscripts*: 103r) in Modern Cornish. Figure 45 shows some examples of aphesis taking place between Middle and Modern Cornish.

Middle Cornish				Modern Cornish	
“alemma”	(*Charter Endorsement*: line 24)	*‘hence’*	>	*“lebah”*	(*Gwavas Manuscripts*: 102v)
“aseth”	(*Pascon Agan Arluth*: stanza 52)	*‘seat’*	>	*“seath”*	(*Gwreans an Bys*: line 65)
“avel”	(*Pascon Agan Arluth*: stanza 114)	*‘like’*	>	*“vel”*	(*William Bodinar’s Letter*)
“dynar”	(*Pascon Agan Arluth*: stanza 36)	*‘penny’*	>	*“in ar”*	(Boorde)
“egerys”	(*Pascon Agan Arluth*: stanza 210)	*‘opened’*	>	*“geres”*	(*Gwavas Manuscripts*: 99v)
“eseza”	(*Pascon Agan Arluth*: stanza 13)	*‘to sit’*	>	*“zetha”*	(*Gwavas Manuscripts*: 103r)
“omscumvnys”	(*Pascon Agan Arluth*: stanza 17)	*‘curse’*	>	*“skemynys”*	(*Gwreans an Bys*: line 213)

Figure 45 Aphesis between Middle and Modern Cornish

Syncope refers to the loss of a medial segment. Thus Middle Cornish “gorthell” (*Gwreans an Bys*: line 2256), ‘a ship’, becomes “goral” (*Gwavas Manuscripts*: 103r) in Modern Cornish. When a vowel is the subject of medial elision, this is referred to as synaeresis. Thus Middle Cornish “omscumvnys” (*Pascon Agan Arluth*: stanza 17), ‘cursed’, becomes “omskemnys” (*Gwreans an Bys*: line 1211) in Modern Cornish. Figure 46 shows some examples of syncope that take place between Middle and Modern Cornish.

Middle Cornish				Modern Cornish	
“beghan”	(*Pascon Agan Arluth*: stanza 166)	‘small’	>	“bean”	(*Gwreans an Bys*: line 117)
“gorthell”	(*Gwreans an Bys*: line 2256)	‘a ship’	>	“goral”	(*Gwavas Manuscripts*: 103r)
“mernas”	(*Pascon Agan Arluth*: stanza 82)	‘unless’	>	“menas”	(*Gwreans an Bys*: line 134)
“mowes”	(*Charter Endorsement*: line 6)	‘girl’	>	“moes”	(Chirgwin)
“mygtern”	(*Pascon Agan Arluth*: stanza 102)	‘king’	>	“matern”	(*Gwavas Manuscripts*: 104r)
“omscumvnys”	(*Pascon Agan Arluth*: stanza 17)	‘cursed’	>	“omskemnys”	(*Gwreans an Bys*: line 1211)
“ordnys”	(*Pascon Agan Arluth*: stanza 151)	‘ordered’	>	“ornys”	(*Gwreans an Bys*: line 1236)
“wolsowas”	(*Pascon Agan Arluth*: stanza 1)	‘heard’	>	“gazowaz”	(*Gwavas Manuscripts*: 101r)
“yn weth”	(*Pascon Agan Arluth*: stanza 136)	‘also’	>	“aweeth”	(*Gwavas Manuscripts*: 104v)

Figure 46 Syncope between Middle and Modern Cornish

Apocope refers to the loss of a final segment. Thus Middle Cornish “dallaʒ” (*Charter Endorsement*: line 25), ‘a start’, becomes “dalla” (*Gwavas Manuscripts*: 103r) in Modern Cornish. Figure 47 shows some examples of apocope that take place between Middle and Modern Cornish.

Middle Cornish				Modern Cornish	
“blyzen”	(*Pascon Agan Arluth*: stanza 228)	‘a year’	>	“bletha”	(*Tonkin Manuscripts B*: 207c)
“dallaz”	(*Charter Endorsement*: line 25)	‘a start’	>	“dalla”	(*Gwavas Manuscripts*: 103r)
“deweth”	(*Pascon Agan Arluth*: stanza 236)	‘an end’	>	“duah”	(*Gwavas Manuscripts*: 101v)
“enaff”	(*Pascon Agan Arluth*: stanza 212)	‘a soul’	>	“ena”	(*Gwreans an Bys*: line 1048)
“flog”	(*Charter Endorsement*: line 21)	‘a child’	>	“flo”	(*Gwavas Manuscripts*: 104v)
“forth”	(*Pascon Agan Arluth*: stanza 15)	‘a road’	>	“vor”	(*Gwavas Manuscripts*: 101v)
“gans”	(*Pascon Agan Arluth*: stanza 234)	‘with’	>	“gan”	(*Gwavas Manuscripts*: 102r)
“kerth”	(*Gwreans an Bys*: line 1381)	‘a walk’	>	“carr”	(*Gwavas Manuscripts*: 105r)
“lowarth”	(*Pascon Agan Arluth*: stanza 140)	‘a garden’	>	“looar”	(*Gwavas Manuscripts*: 99v)
“molloz”	(*Pascon Agan Arluth*: stanza 66)	‘a curse’	>	“mola”	(*Tonkin Manuscripts B*: 207c)
“warbarth”	(*Pascon Agan Arluth*: stanza 127)	‘together’	>	“ware bar”	(*William Bodinar’s Letter*)

Figure 47 Apocope between Middle and Modern Cornish

Diachronic mutation refers to the replacement of a grapheme or cluster of graphemes with another grapheme or cluster of graphemes over a period of time. Two types of diachronic mutation are frequently attested between the Middle and Modern Cornish periods: diphthongisation and pre-occlusion.

Diphthongisation is the replacement of a single vowel grapheme by a diphthong or pair of vowel graphemes. Diphthongisation is quite common

between the Middle and Modern Cornish periods. For example Middle Cornish <E> is frequently found in Modern Cornish as <EA>. Thus the Cornish word for ‘man’ is found in Middle Cornish written “den” (*Pascon Agan Arluth*: stanza 8) and in Modern Cornish written “dean” (*Gwreans an Bys*: line 239). Middle Cornish <O> is frequently found in Modern Cornish as <OA>. Thus the Cornish word meaning ‘bad’ is found in Middle Cornish written “drok” (*Pascon Agan Arluth*: stanza 21) and in Modern Cornish written “droag” (*Gwavas Manuscripts*: 99v). Middle Cornish <U> is frequently found in Modern Cornish as <UE>. Thus the Cornish word meaning ‘wise’ is found in Middle Cornish written “fur” (*Pascon Agan Arluth*: stanza 191) and in Modern Cornish written “fuer” (*Gwreans an Bys*: line 786).

In Middle Cornish <M> and <N> frequently become pre-occluded in Modern Cornish and are attested <BM> and <DN>. In the latest stage of this process, the nasal consonants <M> and <N> may be lost altogether. Thus the Cornish word for ‘this’ is attested in Middle Cornish as “hemma” (*Pascon Agan Arluth*: stanza 86) and is attested in Modern Cornish as “hebma” (*Gwreans an Bys*: line 2500) and as “eba” (*Gwavas Manuscripts*: 101r). The Cornish word for ‘that’ is attested in Middle Cornish as “henna” (*Pascon Agan Arluth*: stanza 5) and is attested in Modern Cornish as “hedna” (*Gwreans an Bys*: line 2448) and as “hedda” (*Gwavas Manuscripts*: 101v).

4.2 The entry-form

The most important part of the lemma is the entry form or head word which

begins the entry and determines that entry's place in the word list. Either a base form or an oblique form may serve as the entry form. The base form is the part of the paradigm that represents the lexeme. The part of the paradigm used to serve as the base form is normally determined by tradition. If at all possible, some indication of the declension or conjugation is evident from the part of the paradigm chosen for the base form. The canonical form is the form chosen by the lexicographer from among the various attested spellings of the base form. Either prescriptive or normative principles may be used when selecting the canonical form. The prescriptive principle is based on posited, authoritative norms. The normative principle, on the other hand, draws on regular usage as attested in a corpus to establish norms.

According to Béjoint (1994: 17-18),

> Every single paragraph that constitutes an entry in a dictionary is headed by a short graphical sequence, the entry-form, which is generally - but not necessarily - the object of the information contained in the entry. In the prototypical dictionary, this sequence is usually a word, in the sense of 'any interrupted sequence of graphemes that is commonly felt to correspond to a concept'. In many modern dictionaries, some entries are also headed by morphemes, mostly prefixes and suffixes. Dictionaries of idioms may have longer strings of words, but there is one word in each string which is used as the classifying unit.

According to Landau (1989:76) "most headwords, with the exception of cross-references and names, are canonical forms". In the case of a cross-reference it may be a variant of the base form or, especially if the paradigm is irregular, an oblique form.

Zgusta (1971: 119) does not distinguish between the terms 'canonical form' , 'basic form' and 'entry-form', and uses the term 'canonical form' for all three.

However I would like to distinguish these terms. I shall use the term ‘entry form‘ to refer to the form that begins a dictionary entry. I shall use the term ‘base form‘ to refer to the part of the paradigm that is used to represent the lexeme. Sometimes the base form is found in a number of different spellings; the ‘canonical form’ then is the preferred or chosen spelling of the base form. Figure 48 shows the entry form system. An entry form my be either a base form or an oblique form. A base form may be the canonical form or it may be a variant spelling of the canonical form.

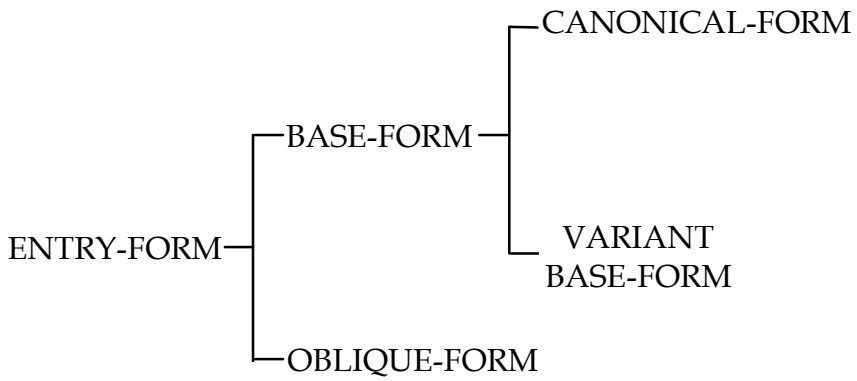


Figure 48 The entry form system

4.2.1 The base form

In spite of the many types of lexical unit that may represent lemmata, separate entries for each of the grammatical words that comprise one lexeme are not usual in monolingual dictionaries, only one grammatical word, the ‘citation’ or ‘base form‘, being entered. The base form is related to its oblique forms (i.e. the other grammatical words that belong to the lexeme) by information concerning inflection. Usually the user needs to know the base form that corresponds to a particular grammatical word that has been encountered, in

order to be able to look it up (Mugdan 1991: 518).

Tradition is usually the deciding factor in determining which part of the paradigm is to be used as the base form. In many languages, for example, it is common practice to use the nominative singular as the base form of nouns. Nevertheless this need not be the rule for all languages that have similar inflections. Different forms are sometimes employed as the base form. The choice of which part of the paradigm is to be used as the base form should be determined by its optimal suitability as a starting point from which to derive the rest of the paradigm (Zgusta 1971: 120).

The paradigm of the Cornish noun has five forms: singular, collective, dual, singulative and plural. Cornish lexicographical tradition has been somewhat inconsistent in choosing the base form of the nominal paradigm; singular, collective and singulative forms are all used as base forms in Cornish dictionaries. Since the dual, singulative and dual forms are all usually formed by the addition of affixes, the singular or collective form is possibly the best choice for the base form. Some nouns, however, have both singular and collective forms. For example, the singular form *davas* (‘a sheep’) also has a collective form *deves* (‘a flock of sheep’). Usually singular forms are distinguished from their collective forms by vowel affection. Since this results in singular and collective forms appearing at some distance from one another in the word list, it possibly best if they are both given entry status and cross referenced to one another. The declension of the noun is not, however, transparent from the uninflected singular or collective form. It is, therefore, necessary to indicate, somewhere in the lemma, the oblique forms of

nouns.

According to Martin (1967: 157),

> One of the traditions of Western lexicography is to use the so-called ‘infinitive‘ form for both the entry heading and the translation of verbs.... In many parts of the world, verbs are usually entered under the plain present (or non-past) form, and it is misleading to translate such headings with English ‘to’ + constructions. Japanese suru does not mean ‘to do’; it means ‘(someone) does’ or ‘will do.’ The one advantage of the ‘to’ + translation is that it clearly marks the word as a verb, and in English many verb forms are homonymous with nouns....

In the case of Cornish, the verbal noun is traditionally chosen as the base form of the verbal paradigm. This has the advantage that the verbal noun suffix indicates to which conjugation the verb belongs.

The paradigm of the Cornish adjective has three forms: positive, comparative and superlative. The comparative and superlative forms are sometimes formed by inflection and sometimes periphrastically. It is the positive form of the adjective which is traditionally chosen as the base form. It is the non-inflected form of the preposition that traditionally serves as its base form. Inflected pronominal prepositions may be cross referenced to their base form. Since so few of the cardinal numbers are inflected, traditionally all inflected forms of cardinal numbers are treated as entry forms.

The base form is, thus, arrived at by a process of deinflection and separation of homonyms. A number of problems are encountered in this process. Some items consist of several distinguishable word forms; these require special treatment. Zgusta (1971: 287) maintains that

> In languages with numerous and regular derivations ... , it is possible to construct very rich and extensive nests. In that case there are two particular difficulties. First, such nests can disagree rather considerably with the alphabetical order. ... The second difficulty is caused by words the morphemic status of which is unclear, whether only to the general user or to the savant himself.

Zgusta (1971: 289) is of the opinion that the practical disadvantages of nests are greater than their practical advantages. He points out that large dictionaries make little use of them but they become necessary, however, when limitations of space are pressing. He also notes that some dictionaries make use of nests for pedagogic or descriptive purposes.

It may help if, in the language being recorded, the groups nested are frequent in order that the dictionary user may become accustomed to the types of nests employed by the lexicographer In the case of a nest conflated from entries whose entry-words are not derivations from the first word, each must have its own statements of meaning. However, the lexicographer will have to take care with this type of nest since the single entries may tend to develop a polysemy of their own and more variation may be displayed by individual members of the nest (Zgusta 1971: 285).

Zgusta (1971: 286) suggests that when employing abbreviations in nests, morphemic boundaries should always be respected. Furthermore, nests should not be based on a graphemic coincidence that lacks genuine morphemic identity.

There may be a number of spelling variants for a given item. Thus the Cornish word meaning ‘now’ is found variously spelled as *lebben*, *leben*, *lebmyn*, *lemman*, *lemmen*, *lemmyn*, *lemyn*, *lymmyn* or *lymyn*.

Kromann, Riiber and Rosbach (1991: 2723) note that in bilingual dictionaries, lemmatisation of lexical units is approached differently depending on whether the dictionary is active or passive. Orthographic variants must all be listed in a passive dictionary, but in an active dictionary, one form will suffice. Schnorr (1991: 2816) notes that with regard to homonym disambiguation, any decision concerning the number of base forms may depend on whether the dictionary is intended for encoding or decoding. A dictionary intended to serve the user for translation from the native to the foreign language, or encoding, is termed an active dictionary, and one intended for translation from a foreign to the native language, or decoding, a passive dictionary. Since a decoding user may experience difficulty knowing the base form, a passive dictionary will need to list many non-base forms

There may be no uninflected base form. A case in point is a Cornish verb, usually translated as 'behold' or 'see', and that is only found in the imperative mood: *ot*, *otte*, *otta* or *yta*. There is, thus, no verbal noun to act as its base form. Sometimes a base form is not attested in the corpus but is assumed to exist in the language system and is reconstructed according to the rules of morphology by the lexicographer. An example is the Cornish ARVA, 'to arm', which is not attested in the corpus, and which, according to George (GKK), has been deduced from the past participle *ervys*, 'armed'. The head word, **arva**, first appears in Morton Nance and Smith's *An English-Cornish Dictionary* (ECD2) of 1934. Several later dictionaries include the reconstructed base form, **arva** (NCED; ECD3; CED; GKK; NSCD).

For certain nouns, the singular base form is not usual or possible. Nouns

which occur only in the plural are known as pluralia tantum and are usually represented by their plural form. Examples of Cornish pluralia tantum include BRUDNYAN (‘groats’), FYANASOW (‘grief’), and GARTHOU (‘ox-goad’). It is necessary to indicate in the lemma that the form is a plurale tantum.

For many languages, feminine forms by convention are treated under masculine forms. However it is sometimes felt that certain feminine forms need to be treated as base forms. Inconsistencies then arise, leaving the lexicographer prone to accusations of sexism (Schnorr 1991: 2813*ff.*). One solution might be to use the masculine form as the base form only when the feminine form consists of the masculine form plus a feminine suffix. Figure 49 shows examples of feminine forms derived from their masculine counterparts by the addition of the suffix –ES. It can be seen that the feminine form *mygternes* is straightforwardly derived from its masculine counterpart *mygtern* by the addition of *-es*. But with many of the other masculine-feminine pairs there is some difference in the stem. This difference may be due to morphological alternation or simply to the free variation in spelling that is so prevalent in the corpus. It is recommended, therefore, that, in the case of Cornish, both masculine and feminine attested forms be used as base forms.

Masculine	Feminine
"du" (*Pascon Agan Arluth*: stanza 24; *Passio Domini*: line 326) "due" (*Origo Mundi*: line 73) "dew" (*Passio Domini*: line 49)	'god'	"dues" (*Origo Mundi*: line 155)	'goddess'
"kentrevek" (*Origo Mundi*: line 2231)	'neighbour'	"kentrevoges" (*Beunans Meriasek*: line 1551)	'neighbour'
"cowyth" (*Origo Mundi*: line 95)	'companion'	"cowethes" (*Origo Mundi*: line 92)	'companion'
"mester" (*Pascon Agan Arluth*: stanza 60)	'master'	"meystres" (*Charter Endorsement*: line 31)	'mistress'
"maw" (*Passio Domini*: line 1794) "mau" (*Passio Domini*: line 2281)	'boy'	"mowes" (*Passio Domini*: line 1876)	'girl'
"mygtern" (*Pascon Agan Arluth*: stanza 102)	'king'	"mygternes" (*Pascon Agan Arluth*: stanza 226)	'queen'
"pehadur" (*Pascon Agan Arluth*: stanza 8)	'sinner'	"peghadures" (*Passio Domini*: line 491)	'sinner'
"pystryor" (*Passio Domini*: line 1767)	'sorcerer'	"pestryores" (*Origo Mundi*: line 2668)	'sorceress'

Figure 49 Derivation by addition of feminine -ES

By convention, Cornish participles are usually lemmatised under their verbal noun. Thus the past participle "bynyges" (*Passio Domini*: line 230), 'blessed', is derived from the verb "benyga" (*Beunans Meriasek*: line 2176), 'to bless', and can be satisfactorily lemmatised under the verbal noun, **benyga** (NCED). A Cornish past participle need not necessarily be translated by an English past participle. Thus the past participle "fleryys" (*Passio Domini*: line 2739) is derived from the verbal noun FLERYE**,** 'stink', but would be translated as 'stinking' not 'stunk'. Nevertheless, there is no reason why *flerye* should not serve as the base form of "flerys". However, sometimes there is no verbal noun to serve as base form. A case in point is the past participle "dyegrys" (*Beunans Meriasek*: line 3667), 'shocked', which is only attested in the past participle. It is necessary in this situation to use this past participle as the base form.

When affixes are very highly productive, it becomes impossible to provide

Some Cornish lexicographers list onomastic terms separately from the main body of the dictionary. Pryce (ACB) includes a separate "Alphabetical List of the Cornish British Names of the Hundreds, Parishes, and Villages in Cornwall". Gendall (PDMC) includes appendices of "Geographical Names" and of "Personal Names". Alternatively onomastic terms may be listed in the dictionary section in their alphabetical place, as we find in the dictionaries of Morton Nance (NCED, CED) and George (GKK, NSCD).

4.2.2 The canonical form

For Landau (1989: 87), the canonical form serves several different purposes: it denotes the preferred spelling; it denotes the normal printed form of the lexical unit, (i.e., whether capitalised or not; whether considered foreign and italicised or naturalised); in most general dictionaries, it denotes syllabification.

A number of variant spellings of the base form may be attested. In this case, the lexicographer must choose one of these as the preferred or canonical form. Béjoint (1994: 101) points out that even if a dictionary gives orthographic variants of the base form, the lexicographer still has to select his/her preferred form to serve as the canonical form. According to Landau (1989:76), a language has to be standardised if its speakers are to recognise grammatical paradigms as being represented by canonical forms.

> Thus before a dictionary can be written for a language, the language must have developed more or less standard spellings or, in a language with various dialects, have a preferred dialect. Variant spellings and dialectal forms can, of course, be given, and for the larger (and especially the historical) dictionaries should be given; but a single form must be chosen as the canonical one.

(Landau 1989: 76-77)

Choice of the canonical form may be influenced according to whether the lexicographer aims to prescribe what s/he considers to be good usage or to describe usage as it is found in the corpus. Standardisation of orthography may be either prescriptive or normative. If standardisation is prescriptive, then it is based on posited, authoritative norms. Normative standardisation, on the other hand, attempts to establish norms by identifying regular usages as attested in a corpus. Frequency plays an important part in this and authority is supported by examples of usage.

Zgusta (1989: 75) maintains that language change is usually equated with deterioration and the aim of prescriptive dictionaries is to prevent this change by fixing the language. Béjoint (1994: 100-2) identifies two ways in which prescriptive dictionaries indicate preferred usage. On the one hand items may simply be omitted from the macrostructure. Alternatively certain items may have usage labels attached to indicate that they are not recommended.

A good example of the prescriptive approach to Cornish lexicography is George's *Gerlyver Kernewek Kemmyn Meur* (GKK). George completely respells Cornish, basing his orthography on his orthoepy which in turn is based on a putative reconstruction of Middle Cornish phonology (George 1984, 1986). George (GKK: 7) writes, "A prime purpose of this dictionary is to establish **Kernewek Kemmyn** as the standard orthography of Revived Cornish." Several writers (Penglase 1994, Williams 1996, Mills 1999) have demonstrated George's reconstruction of Middle Cornish phonology to be

unsound. Penglase (1994) berates the lack of authenticity in Kernewek Kemmyn resulting out of George's purely conjectural reconstruction of Middle Cornish phonology. Williams (1996) lists 25 ways in which the phonology and spelling of Kernewek Kemmyn are erroneous. Mills (1999) gives numerous examples of inaccuracies in George's data. Since reconstructions of historical Cornish phonology are at best conjectural, it is possible to have several competing phonologies. Thus for the foreseeable future, theories concerning Cornish phonology are likely to remain in a state of flux. Any orthography based on a putative phonology is unlikely, therefore, to remain very stable.

Even if a dictionary is not consciously prescriptive it is likely at least to be normative since, in common with other didactic reference works, it encapsulates a particular linguistic model (Zgusta 1971: 210-211; Zgusta 1980: 8; Rey 1982: 30; Béjoint 1994: 101). Rey (1972) distinguishes between the qualitative norm, which forms the basis for prescription, and the quantitative norm which forms the basis for description. The usage and opinion of those considered to be the finest language users, usually well-known writers and educators, provide the corpus from which the lexicographer infers the qualitative norm. The problem here is in determining which writers should be cited to determine usage. In contrast, statistical frequency derived from a corpus designed to represent the speech community as a whole, provides the quantitative norm. In this manner a form is accepted if its frequency of attestation in the corpus exceeds a certain threshold. The problem, then, is for the lexicographer to determine what that threshold should

be.

4.2.3 Compounds

Zgusta (1971: 131) defines a compound word as "such a word the single parts of which have a lexical meaning of their own, if used alone." Compound words consist of two or more free morphemes (DLP2: 63). Traditionally certain scripts mark the boundaries of words by placing spaces between them. The absence of a space or the use of a hyphen may thus be employed to distinguish a compound from a string of separate items (Zgusta 1971: 132). An example of a Cornish compound is the word DENVYDH which has the English translation equivalent 'nobody'. This item is composed of two elements: DEN, meaning 'man', and VYDH, meaning 'not any'.

When a particular element is used as part of a compound it may undergo a change in its lexical meaning. An element of a compound may be semantically depleted. Frequently it is not possible to comprehend the meaning of a compound from the combination of the meanings of its individual elements. A case in point is the compound "penbronnen" ('fool') (*Resurrexio Domini*: line 2096) which is comprised of the free morphemes PEN ('head') and BRONNEN ('rush' *botanical*). The meaning of "penbronnen" is, thus, not transparent from the morphemes of which it is comprised.

Sometimes single parts of a compound have a different form from that used in isolation. The fact that a compound exists may sometimes be obscured when the individual parts of a compound are very changed in there spelling. An unknown or very obscure item may serve as a component of a compound

(Zgusta 1971: 132-133). A case in point is the compound “pednpral” (AB: 52a) which is comprised of the free morphemes, PEDN (‘head’), and an obscure element, PRAL. Morton Nance (NCED: 128) suggests that this second element may be the free morpheme SPRAL (‘clog’ *noun*) with aphesis of initial <S>. George (GKK: 248) simply states that the second morpheme is unidentified.

Differences in the predictability of meaning are of little concern when the purpose of the dictionary is to give a full description of the language. When space is limited, however, the lexicographer may decide to omit those compounds whose meaning is transparent from the constituent elements (Zgusta 1971: 134-5).

In languages where compounds are commonplace, new compounds may usually be constructed at will, in much the same way that one constructs sentences. So long as the listener understands the “rules of coinage”, the speaker may invent new expressions and language is, thus, creative. A compound may, consequently, take the character of a nonce-form created for the occasion of the utterance. The lexicographer should, therefore, not assume that, because an item may be morphologically and orthographically characterised as a compound, it need be treated as more than a combination of single items. Two criteria distinguish compounds: unity of their designative meaning and stability as indicated by their frequency of recurrence (Zgusta 1971: 135-6).

4.3 Alphabetisation

Nowadays the alphabetical arrangement of entries is part of the dictionary's social image (Rey-Debove 1971: 21). In common with the catalogue and the directory, the very genre of the dictionary is associated with the convention of alphabetical arrangement (Malkiel 1975: 17). Lexicographers in Europe decided to begin with the leftmost letter when they first started to arrange words alphabetically. To do this was logical, although they might have decided to classify according to final letters. Initially only the first letter of every entry word was used. Gwavas's Cornish-English glossary (*Gwavas Manuscripts*: 119v-125r) written early in the 18th century is an example of a Cornish word list sorted alphabetically under the first letter of the entry word only. The larger word-lists of later dictionaries required that words be classified within each letter by their second letter, then the third, *etc.*. The word list of Hals' "Lhadymer ay Kernou" (*Gwavas Manuscripts*: 59r-78v), compiled circa 1700 is sorted under the first 2 letters of each entry word. Thereafter Cornish dictionaries tend to employ full alphabetical sorting. Languages such as French, German, Swedish and Czech required small adjustments to be made with regard to diacritics (McArthur 1986: 76; Zgusta 1989; Béjoint 1994: 14).

There are a number of advantages with the alphabetical arrangement of entries. Word by word, paragraph by paragraph division of data focuses the dictionary's subject matter within easily digestible texts (Béjoint 1994: 16). For the dictionary user, alphabetical order is considered to be the easiest and fastest system (Zgusta 1971: 282; Rey 1977: 20-21). It gives the user the

impression that both the universe and the lexicon are exhausted by the dictionary and that the dictionary represents harmony, the totality, the immobile order of things (Rey 1970: 14). Ideally, all entries should be as accessible as possible; irregular oblique forms should, if possible be given a full entry or failing that a cross-reference to the main entry (Householder 1975: 279).

Alphabetisation, however, has its critics, in particular among structuralist linguists. If the dictionary is intended as model of the mental lexicon, or if the dictionary is to be used onomasiologically, to find ideas rather than forms, then it may be argued that an alphabetical arrangement is not ideal. Makkai (1980: 127) maintains that, traditional dictionaries fail to adequately represent the associative groupings of lexemes as a result of their reliance on alphabetisation. In practice, however, an alphabetical index is usually supplied with semantically arranged reference works, in order to facilitate consultation. Indeed several, particularly American, versions of *Roget's Thesaurus* are entirely alphabetically arranged (Béjoint 1994: 15-16).

Landau (1989: 77) stresses the importance of listing inflected forms as head words with cross-references to the canonical forms, because the canonical forms of the source language may not be known by the user of a bilingual dictionary. This is especially important when the inflections differ markedly from the canonical forms, as *yw* ('is') and *bos* ('be'). Slightly more closely related inflections, such as *(g)wrug* ('made') and *gul* ('make'), or *devedhys* 'came' and *dos* ('come') should also be listed if space permits.

4.3.1 Derived forms

Alphabetical order may, sometimes, be disregarded for the sake of a nest, particularly if the dictionary is small with entries which are easy to survey. Virtually all canonical forms are main entries in an unabridged monolingual dictionary. In shorter dictionaries, however, the canonical forms for many words, chiefly regularly formed adverbs and adjectives but also many nouns, are run on at the end of other entries. A small number of lexicographers assemble only the heads of word families in alphabetical order (Zgusta 1971: 285; Malkiel 1975: 17; Landau 1989: 77-78). Barnhart (1975: 163) suggests that, unless there is a meaning or pronunciation problem, derivatives whose root form is clearly recognisable might be listed as run-ons.

In the case of English lexicography, Osselton (1995: 117) observes that Cawdrey experienced problems of ‘dégroupement‘ and ‘regroupement‘ in his English *A Table Alphabeticall* (TA). He suggests that Cawdrey sacrifices alphabetical order in the interests of getting the base form in first in sequences such as:

> **assigne,** appoint, ordaine
>
> **assignation,** appointment

and

> **captiue,** prisoner
>
> **captiuate,** make subject, or a prisoner.

Cawdrey (TA) also brackets together morphologically related words, such as *criminous* and *criminal* which are in his opinion synonymous (Osselton 1995: 120-24). Osselton (1995: 119-120) notes that even today some dictionaries

prefer to give derived forms main entry status, in strict alphabetical sequence, while others prefer to regroup them under a base form.

Zgusta (1971: 284), defines a ‘nest’ as

> a group of entries which is conflated into one; the conflation is effected almost always by the typographical presentation as a run-on (i.e. the single entry words do not begin at a new line) and very frequently by the abbreviation of the entry words. ... The main purpose of this procedure is to save space .... In a very broad generalization, it can be observed that on the whole, nests containing derivations tend to be more conflated than those which deal with different composed words.

For example, PEGHADOR (‘sinner’) may be run on to PEGH (‘sin’), the presumption being that one will have no difficulty understanding PEGHADOR if one knows the meaning of PEGH and of –ADOR (agentive noun ending), both of which are main entries. If, however, the relationship of the derivative to its stem is less transparent, then a run-on is less successful. Thus it is not such a good idea to locate GONADOR (‘sower’) as a run-on of GONYS (‘work’, ‘service’, ‘cultivation’, ‘tillage’).

A change in the root often throws derivatives out of alphabetical order or makes them difficult to recognise. In this event they may either occur in their proper alphabetical place or be listed as run-ons. This entails a conflict of choice between the dictionary user for spelling and the dictionary user for meaning unless the lexicographer has enough space to include both types of information. Frequently a compromise is made in which derivatives without meaning difficulty but with simple spelling or pronunciation difficulty are listed as run-ons (Barnhart 1975: 164). We see these two approaches to the treatment of derivatives if we compare Morton Nance’s *A New Cornish-*

English Dictionary (NCED) with his *A Cornish English Dictionary* (CED). In the NCED of 1938, the entry for **gava**, 'to forgive', is found on page 61; its derivative, **gyvyans**, 'forgiveness', is found on page 78. In Morton Nance's CED of 1955, **gyvyans** is a run-on under the entry for **gava** on page 39. Morton Nance (CED) then gives a cross reference from **gyvyans** to **gava** on page 47.

In the most rigid type of nest, the meaning of the nested entries is predictable by reference to the meaning of the first entry and indication of categorial difference (Zgusta 1971: 285). Landau (1989: 78-79) points out that, whilst it is common practice to run an adverb on to an adjective, lexicographic practice does not normally allow one to run on an adjective to an adverb even if that adverb is more frequently used than the adjective. He discusses the possibility of using frequency studies to decide whether an item should be a head word. Swanson (1975: 66-7) suggests that extensive lists of derivatives, whose synchronic etymology (derivability) is obvious and simple, may be reduced if a dictionary includes a fairly detailed essay on word-formation. It will still be necessary to include some items, however, because of morphophonemic peculiarities.

Zgusta (1971: 287) maintains that

> In languages with numerous and regular derivations ... , it is possible to construct very rich and extensive nests. In that case there are two particular difficulties. First, such nests can disagree rather considerably with the alphabetical order. ... The second difficulty is caused by words the morphemic status of which is unclear, whether only to the general user or to the savant himself.

Zgusta (1971: 289) is of the opinion that the practical disadvantages of nests

are greater than their practical advantages. He points out that large dictionaries make little use of them but they become necessary, however, when limitations of space are pressing. He also notes that some dictionaries make use of nests for pedagogic or descriptive purposes.

It may help if, in the language being recorded, the groups nested are frequent in order that the dictionary user may become accustomed to the types of nests employed by the lexicographer In the case of a nest conflated from entries whose entry-words are not derivations from the first word, each must have its own statements of meaning. However, the lexicographer will have to take care with this type of nest since the single entries may tend to develop a polysemy of their own and more variation may be displayed by individual members of the nest (Zgusta 1971: 285).

Zgusta (1971: 286) suggests that when employing abbreviations in nests, morphemic boundaries should always be respected. Furthermore, nests should not be based on a graphemic coincidence that lacks genuine morphemic identity.

4.3.2 Compounds and multi-word lexemes

According to Zgusta (1971: 289) it may be difficult to ascertain whether a group of words is really stabilised; the lexicographer may, therefore, be uncertain whether the group should be treated as an entry of its own or as a subentry. Swanson (1975: 65) notes that an orthographic space between constituents may result in an item (such as English *no one*) being overlooked

by the lexicographer and, thus, excluded as an entry.

Thus the user may experience difficulty in identifying the entry form of compounds and multi-word lexemes. One way that modern dictionaries still vary slightly in their use of alphabetical arrangement is in the positioning of compounds and multi-word lexemes. There are three methods of positioning compounds and multi-word lexemes in the word list. The first method is to list compounds, set phrases and idioms under the first identifiable element. Thus a compound is treated in the same manner as any other string of letters. This method results in, for example, the order **pen** (‘head’), **penans** (‘penance’), **pen arth** (‘headland’, ‘promontory’). The second method is to group them together in a single block. Thus compounds are classified immediately after their first word. This results in the classification, **pen** (‘head’), **pen arth** (‘headland’, ‘promontory’), **penans** (‘penance’). According to Landau (1989: 82), letter by letter alphabetisation is more usual than word by word and has the advantage that the dictionary user need not know whether a compound is spelled as one word, as a hyphenated word, or as two words.

With the third method, criteria such as the relative importance of each element, and where the user is most likely first to look, may help the lexicographer to decide on classification. Thus compounds are not always listed in the normal order of their elements. The lexicographer has to decide whether to list LAWEN CATH (‘tom-cat’) under **lawen** (‘entire’, ‘uncastrated’) or under **cath** (‘cat’), HANTER DETH (‘midday’) under **hanter** (‘half’) or **deth** (‘day’).

Landau (1989: 82) notes the difficulty of alphabetising verbal idioms.

> Verbal idioms such as 'have one's eye on' are usually 'run in' at the end of the entry for one of the key words of the phrase, in this instance 'have'. The question of which word is most likely to be sought by the user is one that is sometimes impossible to answer. Should the idiom be placed under the first word, or the most important word? Sometimes the first word is variable, as in 'shed' or 'throw light on'. Sometimes it is not easy to say which word is more important, as in 'hang fire'. Most dictionaries prefer to list idioms under the first word, but exceptions are common. Absolute consistency is purchased at the price of the reader's confusion and frustration.

Benson (1989: 6) describes how the elements of a collocation may be distinguished as 'base' and 'collocator'.

> In verb+noun collocations ... the noun is the base, and the verb is the collocator. In adjective+noun collocations ... the noun is once again the base, and the adjective is the collocator. In adverb+verb collocations ... the verb is the base, and the adverb is the collocator. In adverb + adjective collocations the adjective is the base, and the adverb is the collocator.

By using this scheme some dictionaries indicate collocations under the entry for the base, both for encoding as well as decoding (Benson 1989: 7). Schnorr (1991: 2816) suggests that adjective noun fixed collocations be listed under the noun since nouns are looked up more frequently than other parts of speech. Yet another suggestion is to list a multi-word lexeme under each of its constituents (Householder 1975: 279).

Swanson (1975: 65) is of the opinion that it is unnecessary to sub-enter nominal compounds under their constituents, particularly when the first constituent is statistically or otherwise insignificant.

There have been attempts to determine where users attempt to look up multi-word lexemes (Béjoint 1981; Bogaards 1990; Béjoint 1994: 160-2). According

to Béjoint (1994: 161),

> Dictionary users do seem to expect all multi-word units to have one element that is more important than the others, and they seem to prefer to look them up within the entry for this element. This may be because they feel the need to structure their lexical acquisitions by relating every item to other words that they know. The relations between words, whatever they may be, are probably used as a help to memorization. This would mean that dictionaries like LDOCE, for example, are wrong to enter compounds like artificial insemination according to the beginning of their first element: users know that artificial insemination is 'a kind of insemination'.

Speakers of different first languages, however, seem to have different intuitions regarding which element of a multi-word lexeme is key (Bogaards 1990; Béjoint 1994: 161-2).

Rey-Debove (1971: 20) points out that, since a pair of words are located adjacent to one another in alphabetical order if they begin with the same letter (even the same morpheme), alphabetical order may be a little less arbitrary than it appears. Furthermore there are ways of representing semantic links in an alphabetically arranged macrostructure. Rey-Debove (1989: 932) notes that, whilst prior to the 19th century cross references were rarely used, today dictionaries often cross reference semantically related words. Modern dictionaries, furthermore, frequently employ a system of entries and sub-entries. Some lexicographers, however, are of the opinion that "students derive no commensurate benefit from the hours of time wasted in hunting down words not in their obvious alphabetical place," however scientific it may appear to be grouping etymologically related words together.

Zgusta (1971: 289-90) discusses the relative merits of dealing with multi-word lexemes in sub-entries versus giving them entries of their own. He suggests

that it is easier to deal with multi-word lexemes in sub-entries because that allows easy alphabetical insertion under the second or third word. Generality of meaning and its character of a continuum may be illustrated more precisely in one large entry rather than in individual or brief, separate entries. Furthermore a nest of sub-entries may demonstrate the ramification of the meaning of the entry-word within the set of multi-word lexemes. Nevertheless, Zgusta sees no reason why a multi-word lexeme should not be treated in the same manner as other lexical units if it is stabilised.

4.4 The Historical Development of the Cornish Lemma

Osselton (1995: 7) observes that in the case of English lexicography, the entry in monolingual English dictionaries evolved pretty well into the form which is generally expected today between Cawdrey's *Table Alphabetical* (TA) of 1604 and Johnson's *Dictionary of the English Language* (DEL) of 1755. In the case of Cornish, the evolution came later and may, indeed, have been influenced by English lexicographical practice.

Between the 18th and the 20th centuries the Cornish lemma became increasingly more systematic. In the 18th century, the head word list is comprised of both base and oblique forms, and both mutated and radical forms. The semantic unit represented by the head word may be a vocable, a lexeme or a lexical unit. On the scale of rank, the head word may be a multi-word lexeme, a word or a morpheme. The 19th century lexicographers continued to include both base forms and oblique forms, and mutated and radical forms in the head word list. In the 20th century, head words are

generally in the base form; oblique forms tend to only appear as head words when they are not found in the base form; irregular oblique forms are usually cross-referenced to their canonical form. It is also in the 20th century that we see the appearance in Cornish dictionaries of appendices, containing tables of mutations, and paradigms of verbs, pronouns and prepositions.

Manuscript vocabularies of the 18th and 19th centuries tend to have their head word lists sorted alphabetically only by the first one or two letters of the head word. All printed Cornish dictionaries, on the other hand, have their head word lists sorted completely alphabetically. In the 18th century, the usual practice is to conflate <I> and <J>, and <U> and <V> for the purpose of sorting.

The inclusion of variant base forms after the head word is found in the 18th century and continues in the 19th and 20th centuries. Part-of-speech fields first start to appear in the 19th century and have appeared in all Cornish dictionaries since. In the 20th century, we also see fields for mutation, pronunciation and various sorts of etymological information appearing in Cornish dictionaries. Attempts to standardise the spelling of head words appear in the 19th century. Standardisation is one of the main issues throughout the 20th century with a number of competing standards of orthography emerging.

William Hals' "Lhadymer ay Kernou" (*Gwavas Manuscripts*: 59r-78v) was compiled sometime around the year 1700. Each page is divided into three columns. Entries are sorted under the first two letters of the head word only. The head word list runs from A to CLUID. The form of the head word may be

base or oblique and may be mutated or radical. Thus, listed as head words, we find the head word **Bease** (‘a finger’); we find **Bes** (‘praying’) an oblique form of the verb PYSY; and we find **Ben** (‘a head or chief’) a mutated form of PEN. Variant spellings are sometimes listed after the head word in its canonical form. Thus the head word **Bew** is followed by the variant spellings *Bewe* and *Bewn* and its mutated form *Vewn* (see Figure 50). The head word list includes many onomastic terms as well as a great many Latin, Greek and Hebrew terms which strictly have no place in a work that purports to deal essentially with Cornish. This led Pryce (ACB: iv) to criticise Hals’ work as “a most strange hodge-podge of Hebrew, Greek etc. and British words”.

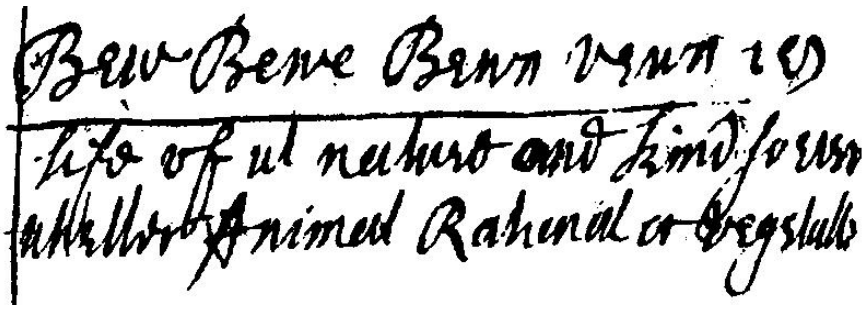


Figure 50 Hals’ Lhadymer ay Kernou (LK)

Lhuyd’s *Geirlyfr Kyrnweig* (GK), compiled sometime after his visit to Cornwall in 1700, is arranged, with a few exceptions, fully alphabetically. Head words in the *Geirlyfr Kyrnweig* are usually single-word base forms.

Gwavas’s Cornish-English glossary (*Gwavas Manuscripts*: 119v-125r) was written early in the 18th century. It is sorted alphabetically under the first letter of the head word only. The head word list runs from A to OZE and confines itself to Cornish language items. Onomastic terms are not included. The head

word list includes several compounds and multi-word expressions. Thus we find the compound, **Bara Sogal** (‘Rye Bread’), and the phrase **fat la er a why o keel** (‘how do you do’) both listed as head words. The form of the head word may be base or oblique and may be mutated or radical. Thus listed as head words we find the base form **Bara** (‘bread’); we find **Cothez** (‘fallen’) an oblique form of the verb **cotha**; and we find **bregaothys** (‘preaching’) a mutated form of **pregaothys**. No variant spellings are listed after the head word. There are no entries beginning with the letter <I> so it is not possible to determine whether Gwavas treats <I> and <J> as separate letters for the purpose of ordering the head word list. Nor do we know if Gwavas would conflate <U> and <V> since the head word list ends at “OZE”. Where it is necessary to clarify some aspect of pronunciation, Gwavas has used colons to separate the syllables of some of the head words (see Figure 51). For example, the syllable marking for the item, **Ac:he:son**, shows that the first consonant is a velar stop [k] not a post-dental affricate [tʃ]. This fact concerning the pronunciation of ACHESON does not seem to have been noted by later Cornish lexicographers (e.g. NCED, CED; GKK, NSCD), though it is confirmed by Lhuyd (AB: 240b) who spells the word, with aphesis, “keyson”.

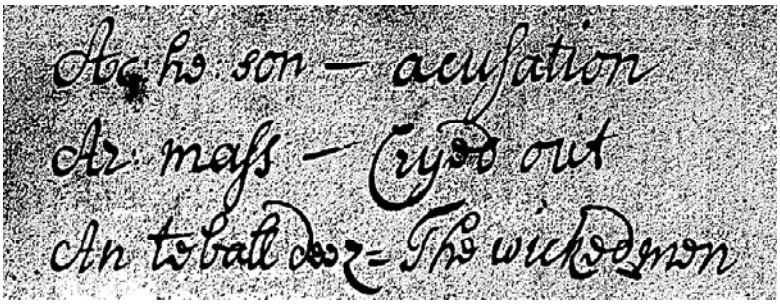


Figure 51 Gwavas’ vocabulary

Borlase's manuscripts (*Mems. of the Cornish Tongue*) contain a handwritten vocabulary from which the published version, "Vocabulary of the Cornu-British Language" (VCBL), was prepared. In the manuscript version, each page is divided into three columns. Entries are sorted under the first two letters of the head word only. In the published version the alphabetical sorting is complete. For the purposes of sorting, <I> and <J> are conflated, and <U> and <V> are conflated. However <I> and <U> are used where they are presumed to be vocalic, and <J> and <V> are used where they are presumed to be consonantal.

The form of the head word may be base or oblique and may be mutated or radical. Thus, listed as head words, we find the base forms **Bealtine** ('fires lighted to Belus') and **Bedh** ('grave'); we find **Be**, **Beazen**, **Beazez**, **Bedh**, **Bedhav**, **Bedhez** and **Bedhon** - oblique forms of BOS ('to be') also listed as head words; and we find **beb** ('every one') a mutated form of PUB (see Figure 52).

Be, *he hath been.*
Bealtine, *Fires lighted to Belus.* Ir.
N. B. The Corniſh for Fire is Tan; but to tine, or light a a Fire, is ſtill uſed in Cornwall, unde Bartine, *the fiery top*, i. e. *the hill of Fires.*
Beazen, beaze, beazenz, *we, ye, they had been.* V.
Beazez, *thou hadſt been*; beaze, *he had been.*
Beb, *every one*; pub, id.
Beeh, *a Voyage.* Ar.
Beehye, *to thruſt.* V.
Bederow, *Prayers.*
Bedh, & Bez, *be thou.*
Bedh, *a Grave*; pl. bedhiow, bethow, id.
Bedhav, *I will be*; bedhi, byd, *thou, he will be.*
Bedhez, boez, biz, *let it be.*
Bedhon, bedhoh, bedhanz, *we,*

Figure 52 VCBL, Be - Bedhon

The semantic unit represented by the lemma may be a vocable, a lexeme or one single sense of a lexeme. Borlase usually gives homographs separate entries. Thus the form *Da* is given 3 separate entries since 3 distinct lexemes are evident (see Figure 53).

DA (dha, & dah, id.), *good.*
Da, *thy*; tha, id.
Da, *a Doe.* Cott.

Figure 53 VCBL, Da

Occasionally, however, the entry represents a vocable. Thus two

lexemes are conflated under the entry for **Côr** (see Figure 54); the two translation equivalents, ‘ale’ and ‘manner’, given in the entry are completely unrelated.

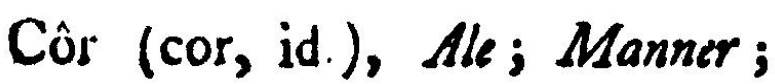

Côr (cor, id.), *Ale*; *Manner*;

Figure 54 VCBL, Côr

Occasionally, Borlase gives a separate entry to each sense of a lexeme. Thus the lexeme, KORNAT, is given two entries: one for the sense, ‘angle’; and one for the sense, ‘corner’ (see Figure 55). Borlase’s word list is, thus, not a list of lexemes but of forms.


Kornat, *an Angle*. Lh.
Kornat, *a Corner*; *Angulus*. L.

Figure 55 VCBL, Kornat

The unit of rank represented by the lemma may be a multi-word unit, a single-word unit or a morpheme. Thus we find an entry for the single word, **Ban** - ‘up’; the multi-word unit, **Ban a sevy** - ‘up he stood’; and also the suffix **–ik**.

Use of hyphens to denote compounding is common but sometimes the hyphen is omitted. Thus we find **Dama-widen** hyphenated, but **Hernan guidn** unhyphenated.

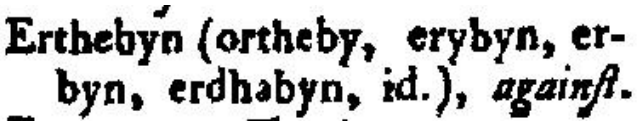
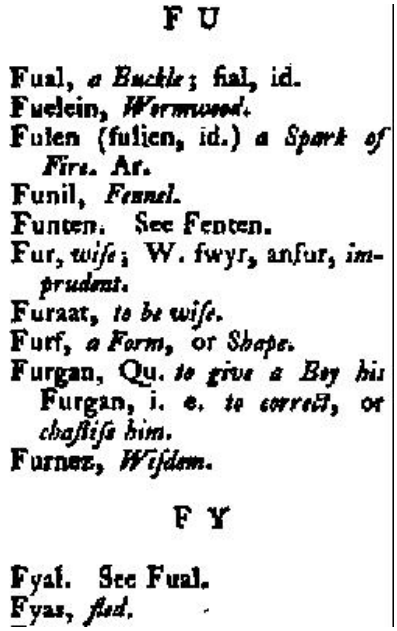

Erthebyn (ortheby, erybyn, er-
byn, erdhabyn, id.), *againſt*.

Figure 56 VCBL, Erthebyn

Variant spellings are frequently listed separately in the word list. Thus **Me**, **Mi** and **My** are each given separate entries. Sometimes Borlase lists variant forms after the head word. Thus the head word **Erthebyn** is followed by the variant forms, *ortheby*, *erybyn*, *erbyn*, *erdhabyn* (see Figure 56). Some entries are merely cross-references to a preferred spelling. Thus the head word **Fyal** cross-refers to the entry under **Fual** (see Figure 57).


F U

Fual, *a Buckle*; fial, id.
Fuelein, *Wormwood*.
Fulen (fulien, id.) *a Spark of Fire*. Ar.
Funil, *Fennel*.
Funten. See Fenten.
Fur, *wiſe*; W. fwyr, anfur, *imprudent*.
Furaat, *to be wiſe*.
Furf, *a Form*, or *Shape*.
Furgan, Qu. *to give a Boy his* Furgan, i. e. *to correct*, or *chaſtiſe him*.
Furnez, *Wiſdom*.

F Y

Fyal. See Fual.
Fyas, *fled*.

Figure 57 VCBL, Fual - Fyas

Lemmatisation of the "Cornish-English Vocabulary" in Pryce's *Archaeologia Cornu-Britannica* (ACB) of 1790 follows much the same principles as Borlase (VCBL). Sorting on the whole is, however, completely alphabetical, though

there are occasional idiosyncrasies. For example the third entry for **Guas** is out of place in the alphabetical sequence (see Figure 58).

GUAR-HEK, *to ride on an horſe.*
Guâs, *mean, one of the commonalty*; guazna, *that fellow; pl.* Guſkâs.
† GUAS, *a cunning, ſubtle fellow.*—*See* BATHOR. By which it ſhould ſignify, *a man of money.*
GUASGA, dho guaſga, *to ſqueeze, preſs, ſtrike*; guaſk, *ſtrike.*
GUASTIA, dho guaſtia, *to conſume, to waſte.*
Guâs, *is alſo, hungry*; guâs decroter thym ya, *an hungry deſire to me there is.*
GUASANAETH, *bondage, ſlavery*; vêz a'n chy guaſanaeth, *out of the houſe of bondage.*

Figure 58 Entry for Guas in ACB

Like Borlase (VCBL), Pryce (ACB) conflates <I> and <J>, and <U> and <V>, for the purpose of sorting the word list. And like Borlase's VCBL, the form of the head word may be base or oblique and may be mutated or radical; the semantic unit represented by the lemma may be a vocable, a lexeme or one single sense of a lexeme; the unit of rank represented by the lemma may be a multi-word unit, a single-word unit or a morpheme. Thus, based on the vocabularies of Borlase (VCBL) and Pryce (ACB), Figure 59 is a system network of lemmatisation of the Cornish word list in the 18th century.

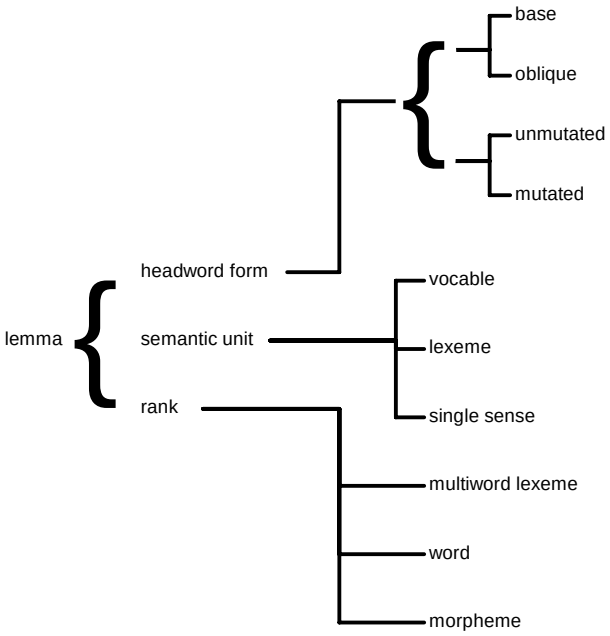


Figure 59 System network of 18th century lemmatisation

In Rogers' "Vocabulary of the Cornish Language" (VCL) of 1861, the head word list is arranged alphabetically under the first two letters of the head word only and includes base forms, oblique forms, and onomastic terms. Variant spellings are sometimes listed after the head word in its canonical form. Rogers (VCL: 5-6) recognises the difficulties that arise from the capricious spelling practices of his sources and writes,

> … the words … are spelt, in reference to their supposed original pronounciation; their [sic] being no standard by which to judge of their accuracy: in fact my principal difficulty has lain in the endeavour to ascertain the correct mode of spelling many of the words according to their primitive construction & original manner of pronounciation. …. For this purpose, I have frequently thought it requisite in order to better understand the genuine meaning to make some alteration in the modern spelling of the words; for many of them appear to me to be so modernized as frequently to convey a very different import from that intended by their original use.

Thus, in Rogers' VCL we find an early attempt towards standardisation

of spelling. It is interesting that Rogers gives "supposed original pronounciation" as a criterion for respelling as this criterion is also used by later lexicographers as a basis for respelling Cornish. It is not clear, however, what grounds Rogers has for supposing an "original pronounciation".

In Williams *Lexicon Cornu-Britannicum* (LCB) of 1865, the lemma consists of 2 fields: the head word and the part-of-speech (see Figure 60).

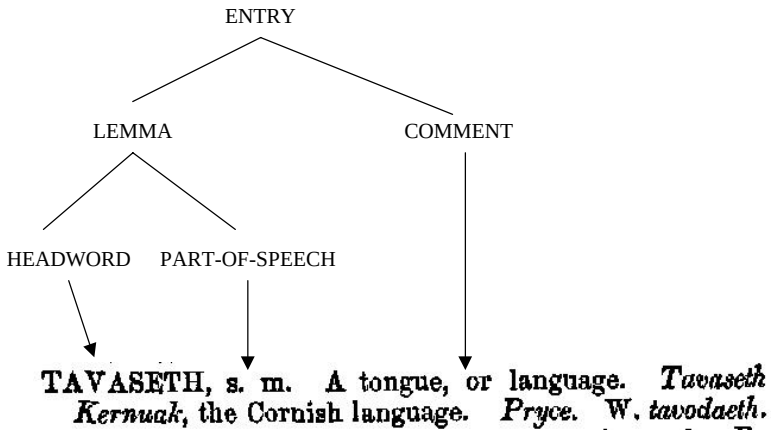


Figure 60 The lemma in LCB

Head words are given in block capitals. Like his predecessors Borlase (VCBL) and Pryce (ACB), Williams (LCB) includes base forms, mutated forms and oblique forms in his head word list. The unit of rank represented by the lemma is mainly the word. However a small number of bound morphemes are also listed; these are all prefixes (e.g. **COV-, DAR-, DAS-, DI-, DIS-, DY-, DYN-, GOR-**, etc.). There are no multi-word units amongst the word list.

There are a small number of ghost words to be found in Williams' (LCB) word list. The base form **GALLY** is not attested in the corpus on which

Williams bases his *Lexicon Cornu-Britannicum*; though it is given by Pryce (ACB). Borlase (VCBL) gives **Galla** as the base form of this lexeme. Williams (LCB) gives another base form **TALY** which is similarly unattested in the corpus; Pryce (ACB) gives **Tyly** as the base form of this lexeme.

Unlike his 18th century predecessors, Borlase (VCBL) and Pryce (ACB), Williams (LCB) does not conflate <I> and <J>, or <U> and <V>. This inevitably involves some respelling if <I> and <U> are taken be vocalic, and <J> and <V> are consonantal. For example, as one of his sources, Williams (LCB) uses Norris' (1859a) transcription of *Origo Mundi* in which we find the form "vhelder"; Williams (LCB) assumes the first phoneme to be vocalic and respells this **UHELDER**. Thus, as with Rogers' VCL, supposed pronunciation is the criterion for respelling.

Respelling in the *Lexicon Cornu-Britannicum* (LCB) is extensive. All letters <K> are respelt as <C>. <TH> is sometimes respelled as <DH>, and <ʒ> is sometimes respelled as <TH> and sometimes <DH>. Williams (LCB) may have used Lhuyd (AB) as a source to distinguish <TH> and <DH> or alternatively might have used analogy with Welsh. Williams (1865), nevertheless, lists no words beginning <DH-> or <TH->; one would, however, expect DHA, DHE and DHI to be included in the word list. At the end of the list of words under <D> is a short section headed "DH" (although there is no corresponding section under <TH>) where Williams (LCB: 130) writes,

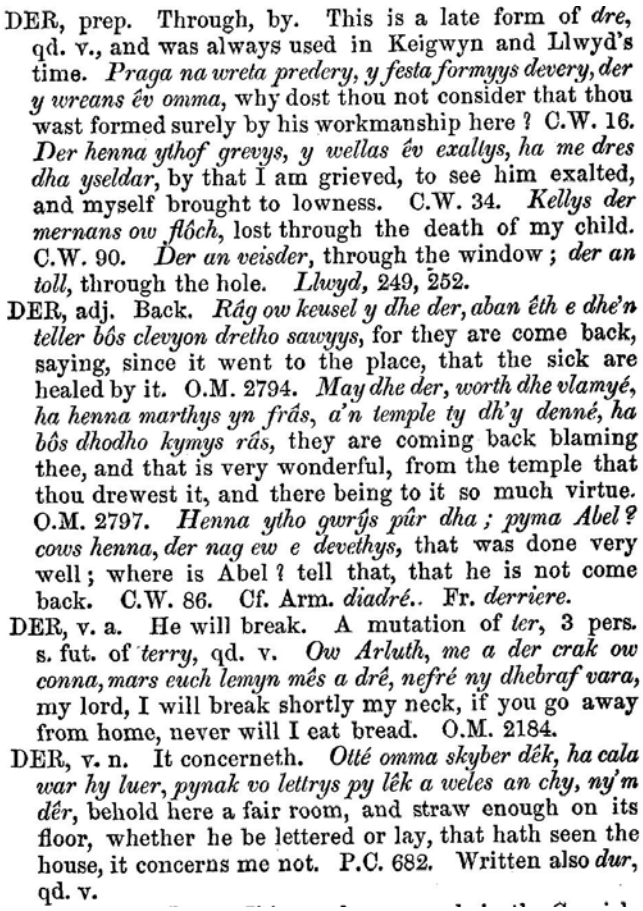
> This is a secondary letter, and is the soft mutation of d, as davas, a sheep; an dhavas, the sheep. … All Cornish words beginning with dh, as dhe, to; dhedhy, to her; dhodho, to him, &c., must be sought for under the primary initial, as de, dedhy, dodho, &c. The Cornish dh is generally written th in the MSS.

Williams (1865) respelling often seems to be sometimes purely capricious. For example, he gives the head word, **HOULSEDHAS**, though his only sources for this item are Lhuyd (AB: 104c) and Pryce (ACB: n.p.), who both spell it "houlzedhas". Williams (LCB) frequently cites "Llwyd" (i.e. Lhuyd) as his source, but does not use Lhuyd's General Alphabet. There are a few items beginning 'DZH-' for which Williams' source is Lhuyd's *Archaeologia Britannica* (AB). Lhuyd (AB) uses 'DZH' to represent the affricate [dʒ]. However, Williams (LCB) lists these under <D> not <J>.

Williams' (LCB) respelling foreshadows the revivalist dictionaries of the 20th century (NCED, CED; GKK). If, however, Williams intended the *Lexicon Cornu-Britannicum* to be used for decoding the published critical editions of classical Cornish texts (Gilbert 1826, 1827; Norris 1859a; Stokes 1863) that he uses as his corpus for the dictionary, then all this respelling only serves to hinder the dictionary user; the more so since Williams gives no explanation of the principles that he has used for respelling items in his word list.

Williams (LCB) treats homographs under separate entries. The part-of-speech field serves as a distinguisher between certain homographs. Thus out of a total 504 sets of homographs in the Williams' (LCB) word list, 391 sets of homographs are distinguished by their part-of-speech. Thus Williams (LCB) includes 4 homographs of *der* amongst his word list which are distinguished by being a preposition, an adjective, a verb active, and a verb neuter (see

Figure 61).



DER, prep. Through, by. This is a late form of *dre*, qd. v., and was always used in Keigwyn and Llwyd's time. *Praga na wreta predery, y festa formyys devery, der y wreans êv omma*, why dost thou not consider that thou wast formed surely by his workmanship here? C.W. 16. *Der henna ythof grevys, y wellas êv exaltys, ha me dres dha yseldar*, by that I am grieved, to see him exalted, and myself brought to lowness. C.W. 34. *Kellys der mernans ow flôch*, lost through the death of my child. C.W. 90. *Der an veisder*, through the window; *der an toll*, through the hole. *Llwyd*, 249, 252.

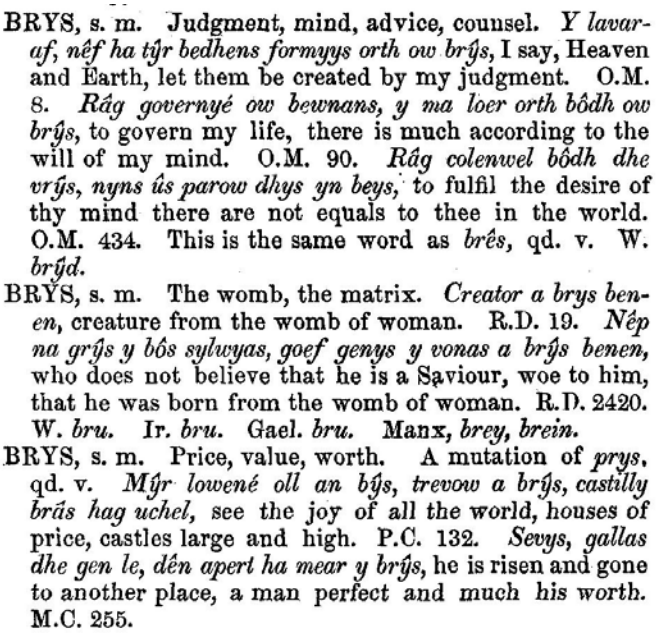
DER, adj. Back. *Râg ow keusel y dhe der, aban êth e dhe'n teller bôs clevyon dretho sawyys*, for they are come back, saying, since it went to the place, that the sick are healed by it. O.M. 2794. *May dhe der, worth dhe vlamyé, ha henna marthys yn frâs, a'n temple ty dh'y denné, ha bôs dhodho kymys râs*, they are coming back blaming thee, and that is very wonderful, from the temple that thou drewest it, and there being to it so much virtue. O.M. 2797. *Henna ytho gwrýs pûr dha; pyma Abel? cows henna, der nag ew e devethys*, that was done very well; where is Abel? tell that, that he is not come back. C.W. 86. Cf. Arm. *diadré*.. Fr. *derriere*.

DER, v. a. He will break. A mutation of *ter*, 3 pers. s. fut. of *terry*, qd. v. *Ow Arluth, me a der crak ow conna, mars euch lemyn mês a drê, nefré ny dhebraf vara*, my lord, I will break shortly my neck, if you go away from home, never will I eat bread. O.M. 2184.

DER, v. n. It concerneth. *Otté omma skyber dêk, ha cala war hy luer, pynak vo lettrys py lêk a weles an chy, ny'm dêr*, behold here a fair room, and straw enough on its floor, whether he be lettered or lay, that hath seen the house, it concerns me not. P.C. 682. Written also *dur*, qd. v.

Figure 61 The homograph *der* in LCB

113 sets of homographs are not distinguished by their part-of-speech. Thus the 3 homographs of *brys* that Williams (LCB) includes are all masculine nouns (see Figure 62).



BRYS, s. m. Judgment, mind, advice, counsel. *Y lavaraf, nêf ha tŷr bedhens formyys orth ow brŷs*, I say, Heaven and Earth, let them be created by my judgment. O.M. 8. *Râg governyé ow bewnans, y ma loer orth bôdh ow brŷs*, to govern my life, there is much according to the will of my mind. O.M. 90. *Râg colenwel bôdh dhe vrŷs, nyns ûs parow dhys yn beys*, to fulfil the desire of thy mind there are not equals to thee in the world. O.M. 434. This is the same word as *brês*, qd. v. W. *brŷd*.

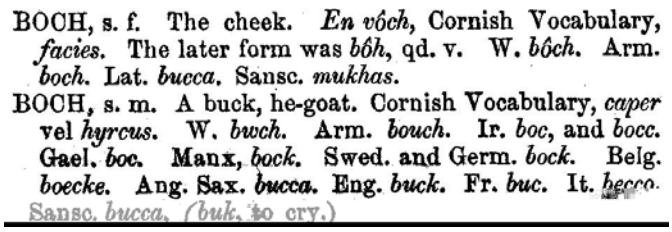
BRŸS, s. m. The womb, the matrix. *Creator a brys benen*, creature from the womb of woman. R.D. 19. *Nép na grŷs y bôs sylwyas, goef genys y vonas a brŷs benen*, who does not believe that he is a Saviour, woe to him, that he was born from the womb of woman. R.D. 2420. W. *bru*. Ir. *bru*. Gael. *bru*. Manx, *brey, brein*.

BRYS, s. m. Price, value, worth. A mutation of *prys*, qd. v. *Mŷr lowené oll an bŷs, trevow a brŷs, castilly brâs hag uchel*, see the joy of all the world, houses of price, castles large and high. P.C. 132. *Sevys, gallas dhe gen le, dên apert ha mear y brŷs*, he is risen and gone to another place, a man perfect and much his worth. M.C. 255.

Figure 62 The homograph *brys* in LCB

The part-of-speech categories used by Williams (LCB) are as follows: adj. (adjective), adv. (adverb), adv. comp. (adverb compounded), art. (article), definite article, conj. (conjunction), conj. pron. (conjunction pronoun), interj. (interjection), num. (number), num. adj. (number adjective), card. num. (cardinal number),pron. (pronoun), pron. adj. / pr. adj. (pronoun adjective), pron. dem. (pronoun demonstrative), pron. pers. (pronoun personal), pron. poss. (pronoun possessive), pron. prep. (pronoun preposition), pron. rel. (pronoun relative), pron. s. / pr. subs. (pronoun substantive), comp. pron. (compounded pronoun), s.m. (substantive masculine), s.f. (substantive feminine), v. (verb), v.a. (verb active), v.imp. (verb imperative), v.irr. (verb irregular), v.n (verb neuter), v.pass. (verb passive), v.subs. (verb substantive), part. (participle). Attribution of gender to nouns distinguishes some pairs of homographs. Thus *boch* has two entries: one as a masculine noun and one as a

feminine noun (see Figure 63).

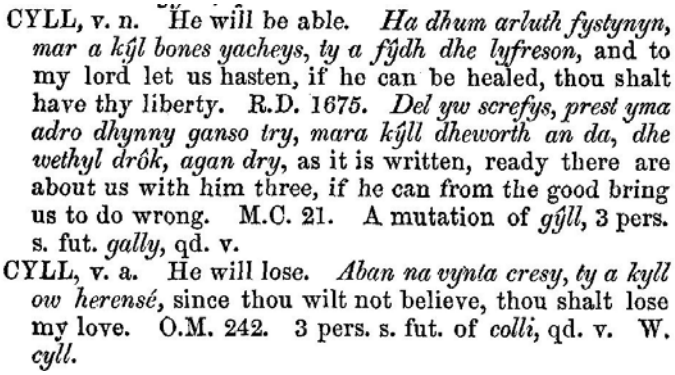


BOCH, s. f. The cheek. *En vôch*, Cornish Vocabulary, *facies*. The later form was *bôh*, qd. v. W. *bôch*. Arm. *boch*. Lat. *bucca*. Sansc. *mukhas*.

BOCH, s. m. A buck, he-goat. Cornish Vocabulary, *caper* vel *hyrcus*. W. *bwch*. Arm. *bouch*. Ir. *boc*, and *bocc*. Gael. *boc*. Manx, *bock*. Swed. and Germ. *bock*. Belg. *boecke*. Ang. Sax. *bucca*. Eng. *buck*. Fr. *buc*. It. *becco*. Sansc. *bucca*, (*buk*, to cry.)

Figure 63 The homograph *boch* in LCB

Williams' (LCB) sub classification of verbs as active, passive, neuter or substantive also serves to disambiguate some homographs. Thus *cyll* has two entries: one as a verb neuter and one as a verb active (see Figure 64).



CYLL, v. n. He will be able. *Ha dhum arluth fystynyn, mar a kŷl bones yacheys, ty a fŷdh dhe lyfreson*, and to my lord let us hasten, if he can be healed, thou shalt have thy liberty. R.D. 1675. *Del yw screfys, prest yma adro dhynny ganso try, mara kŷll dheworth an da, dhe wethyl drôk, agan dry*, as it is written, ready there are about us with him three, if he can from the good bring us to do wrong. M.C. 21. A mutation of *gŷll*, 3 pers. s. fut. *gally*, qd. v.

CYLL, v. a. He will lose. *Aban na vynta cresy, ty a kyll ow herensé*, since thou wilt not believe, thou shalt lose my love. O.M. 242. 3 pers. s. fut. of *colli*, qd. v. W. *cyll*.

Figure 64 The homograph *cyll* in LCB

Where homographs are distinguished by their part-of-speech, Williams (LCB) does not list these in any particular order. Occasionally part-of-speech has been wrongly ascribed by Williams (LCB). Thus **DUETH**, a verb, is wrongly labelled "s.f." (substantive feminine) (see Figure 65).

DUETH, s. f. He came. 3 pers. s. preterite of *dôs*. *Râk whêth bŷth ny dhueth deydh brues*, for yet the day of judgment has not come. R.D. 234. *Namn 'agan dallas golow, pan dhueth an gwâs*, light almost blinded us, when the fellow came. R.D. 303. *Pyw a dhueth a'n beys yn rûdh*, who is it that has come from the earth in red. R.D. 2499. Written also *dûth*, qd. v. W. *daeth*.

Figure 65 DUETH in LCB

In Morton Nance's *A New Cornish-English Dictionary* (NCED) of 1938, the head word list consists of base forms, inflected forms of lexemes which are not found in their base form, and some suffixes and prefixes. Irregular oblique forms are cross-referenced to their canonical form. Appendices are provided with a chart of mutations and the paradigms of verbs, pronouns and prepositions. The lemma may include fields for the head word, the part-of-speech, the etymology, and oblique forms (see Figure 66).

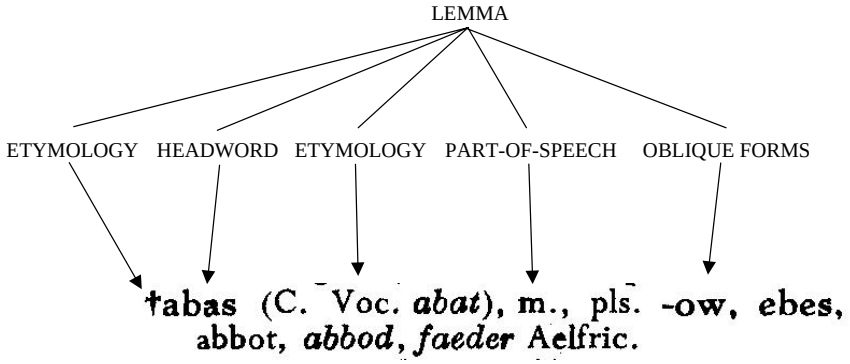


Figure 66 The lemma in NCED

The head word does not necessarily begin the entry. It may be preceded by either a dagger-symbol (†), to indicate that the word is respelt from Old Cornish, or an asterisk (*), to indicate that it is a neologism adapted from Breton or Welsh. The presence of either the dagger-symbol or the asterisk does not effect the normal alphabetical sorting of the word list.

Brackets indicate elements which may undergo elision (see Figure 68).

Most significant about Morton Nance's NCED is the standardisation of Cornish spelling. The Unified Cornish spelling which Morton Nance created is based on the spelling found in the Middle Cornish texts. However, Unified Cornish is not identical to Middle Cornish spelling practice; rather, it is a simplification of Middle Cornish. Frequently words are re-spelt by Morton Nance. The long-tailed-z character <ȝ> found in Middle Cornish is replaced by <TH> or <DH>. Morton Nance also respells words from Old Cornish, Modern Cornish, Cornish dialect and Cornish place-names. In order to fill the gap of lost words, borrowed Breton or Welsh words are re-spelt. Morton Nance's unified spelling has received some criticism. Thomas (1972) complains that Unified spelling has never been explained, in other words there was never any real discussion of the principles on which it was based. Thomas is also critical of the phonological basis of Morton Nance's unified spelling.

Spellings as they are attested in their original form in the corpus and variants are added in brackets, although Lhuyd's (AB) General Alphabet is represented in ordinary type. The Modern Cornish and contracted Middle Cornish forms are given, with reference to which Morton Nance states, "... the form first given being usually preferable, even when it differs from that most usual." Word combinations that are translated by one word in English are hyphenated.

A section on pronunciation is included in the front matter of Morton Nance's

NCED. Within the lemma, a bullet point is placed after a vowel to indicate stress other than on the penultimate syllable. A macron is placed over a vowel to indicate that it is long. A dieresis, <ü>, distinguishes the rounded close front vowel from the rounded close back vowel, which Morton Nance writes <u> (see Figure 67).

a-dhevȳ's, *adj.*, exact, just, right, elegant, complete : see **devȳs** (M.E.).
a-dhewha'ns (*a thyhons* O.M. 2810), *adv.*, immediately : see **dewhans.**
***adhüly,** *vb.*, to adore (W., B.) : replaced by **gordhya.**

Figure 67 Diacritics in NCED

Mutated forms are not given in the head word list. Instead a table of mutations is given in the front matter in order that the dictionary user can find the base forms of items that have undergone initial mutation. Words which cause initial mutation of the immediately following word are marked in the lemma with a superscript numeral to show which state of mutation they cause. This mutation mark also helps to disambiguate homographs. Thus the possessive pronoun, OW (‘my’), which causes third state mutation, is distinguished from the interjection, OW, which does cause mutation, and also from the present participle verbal particle, OW, which causes fourth state mutation (see Figure 68).

ow[4] **(th)-**, *pres. participle verbal particle,* -ing : for uses see APPENDIX VII.
ow[3], *poss. pron.*, my : after prep. or conj. ending in vowel, *'w*; *ow chy ow-honen*, my own house ; see **am**.
ow, *interj.*, ho ! hullo ! : sometimes prefixed to *ot, otta*, as a var. of **awot, awotta.**

Figure 68 Mutation marks in NCED

Figure 69 shows the part-of-speech markers found in Morton Nance's *A New Cornish-English Dictionary* (NCED).

abst.	abstract noun
abstract pl.	abstract plural
adj.	adjective
adj. irreg. comp.	irregular comparative adjective
adv.	adverb
adverbial particle	adverbial particle
art.	article
card. num.	cardinal number
col	collective noun
comp. adj.	comparative adjective
conj.	conjunction
def. vb.	defective verb
dem. pron.	demonstrative pronoun
dual prefix	dual prefix
exclam. verbal particle	exclamative verbal particle
f.	feminine noun
indef. pron.	indefinite pronoun
infixed pron.	infixed pronoun
interj.	interjection
interr. pron.	interrogative pronoun
irreg. vb.	irregular verb
m.	masculine noun
neuter	neuter noun
num.	number
ord. num.	ordinal number
p. pt.	past participle
pl.	plural
poss. pron.	possessive pronoun
prefix,	prefix
prep.	preposition
pron.	pronoun
rel. and interr. verbal particle	relative and interrogative verbal particle
rel. pron.	relative pronoun
suffix,	suffix
suffixed pron.	suffixed pronoun
vb.	verb

Figure 69 Part-of-speech markers in NCED

Irregular verbs are marked "*irreg. vb.*" in the part-of-speech field and their full paradigms are given in an appendix. 21 entries do not contain a part-of-speech field; these are for the following head words: **atta last, brastereth, croadur, dyalar, fortynya, goscor, j'oue, kehesnos, len, lollas, motty, pensogh,**

pocar, potestas, praydha, sampel, to-, trosken, trystys, warn, y praya / y praydha.

It can be seen how the various elements of the lemma distinguish between homographs by examining the entries in Morton Nance's *A New Cornish English Dictionary* (NCED). Figure 70 shows the entries for the homograph, *cuth*. The pronunciation diacritics distinguish **cüth** from **cūth**. The part-of-speech field distinguishes **cüth** the masculine noun from **cüth** the adjective. There is also a second masculine noun **cüth** given as a run-on of **cüth** the adjective. Presumably Morton Nance treats these under the same entry because he considers the part-of-speech distinction between these to be derivational. Furthermore treating this masculine noun as a run-on of the adjective distinguishes it from the other masculine noun homograph of **cüth**. There is also a cross-reference to **cüdha, cüdhy**. **Cüdha** is the verbal derivation of **cüth** the adjective and **cüdhy** is the verbal derivation of **cüth** the masculine noun. Morton Nance's cross-referencing is inconsistent and untidy. He places a cross-reference to **cüdhy** at the end of the entry for **cüth** the masculine noun but he does not place a cross-reference to **cüdha** at the end of the entry for **cüth** the adjective.

cüth, *m.*, sorrow, grief, trouble, travail : *c. ny-gan-gas*, we shall not cease from sorrow, R.D. 2456, lit. sorrow will not leave us ; see **cüdhy**.
cüth, *adj.*, concealed, secret ; *m.*, hiding-place (W., B.).
cüth, see **cüdha**, **cüdhy**.
cūth, *f.*, *pl.* **-ow**, husk, pod, Lh. : *c.-faf*, bean-pod ; *c.-pys*, peascod ; *cuthow*, chaff (D. "cutha", dregs, perhaps confused with **godhas**).

Figure 70 The homograph *cuth* in NCED

In Figure 71, we see the three homographs of the word type *crys*. The second of these is marked with a dagger to show that it is respelled from Old Cornish. The compounds **crys-hok** and **cryspows** are listed as separate entries. **Cryspows** is marked with an asterisk to indicate that it is adapted from Welsh or Breton. The bar diacritics over the letter <ȳ> indicates a long vowel (like 'ee' in English 'seen').

crȳs, *m.*, vigour, vehemency, force, speed : in phrase *gans mur·grys*, *mur a grys*, forcibly, hastily, etc.
†**crȳs** (C. Voc. *kreis*) *m.*, *pl.* **-yow**, shirt, shift, chemise.
crȳs, *m.*, *pl.* **-yow**, shake, shiver, quake : see **dorgrȳs**.
crys-hōk (*krysat* Lh.), *m.*, *pl.* **-ys**, kestrel, "cress-hawk" (E.D.).
***crȳspows**, *f.*, waistcoat : *c. oferyas*, cassock (W.).

Figure 71 The homograph *crys* in NCED *Dictionary*

Morton Nance's *Cornish-English Dictionary* (CED) of 1955 follows more or

less the same principles of lemmatisation as his NCED. Figure 72 shows the entries for the homograph, *cuth*. If one compares this with the entries for *cuth* from Morton Nance's NCED, one can see that there are now more run-ons of **cüth** the masculine noun and that **cüdhy** is now treated as run-on rather than as a cross reference. However inconsistency remains since **covva** and **cüdha** are treated as cross-references rather than run-ons of **cüth** the adjective.

cüth, *m.* sorrow, grief, trouble, travail; **cüdhy,** *vb.* to grieve, make sorry; **cüdhyjyk,** *adj.* sorry, contrite, repentant; ***cudhyjygeth,** *m.* contrition; **cudhyjykhē·,** *vb.* to cause to repent.
cüth, *adj.* hidden, concealed, secret; *m.* hiding, hiding-place: see **covva, cüdha.**
cūth, *f., pl.* **-ow,** husk, pod; **c. fāf,** bean-pod, **c. pȳs,** peascod, pl. used also of chaff, dregs, etc.

Figure 72 The homograph *cuth* in CED

In George's *Gerlyver Kernewek Kemmyn* (GKK) of 1993, the head word list consists of base forms, irregular oblique forms, inflected forms of lexemes which are not found in their base form, suffixes, prefixes and stems. Unlike Morton Nance (NCED), George (GKK) does not include an appendix with paradigms of the verbs. The lemma may include fields for the head word, a disambiguator, authentication, pronunciation, the part-of-speech, and oblique forms (see Figure 73).

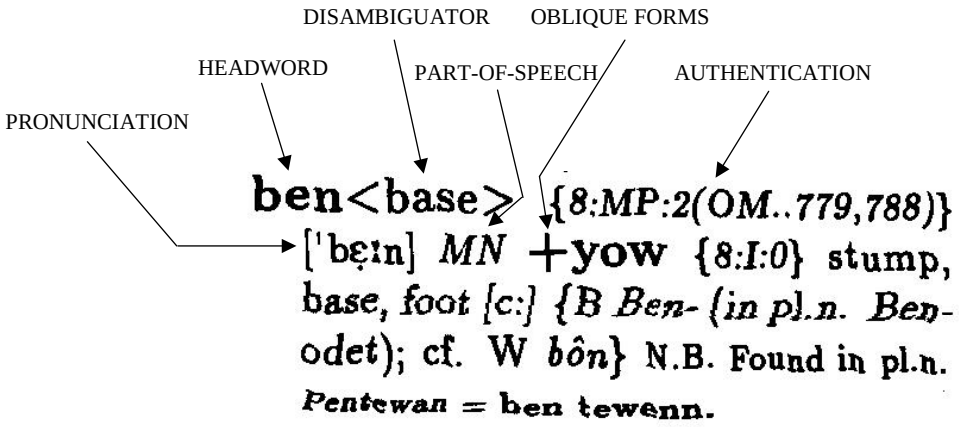


Figure 73 The lemma in GKK

The disambiguator is used to distinguish homographs. In the example in Figure 73, the disambiguator <base> is used to distinguish this item from its homograph **ben** <FN>. The information contained in this disambiguator is redundant since 'base' is included as an English translation equivalent in the comment and the "FN", of **ben** <FN>, is a repetition of the part-of-speech field. In nearly all cases where George (GKK) has used a disambiguator, it is unnecessary. In the few cases where the lemma requires further disambiguation, a conventional genre field label or a number would suffice.

The authentication code deals with the item's etymology. George (GKK: 12) describes its purpose as follows.

> It is important that each word in Revived Cornish be seen as authentic, and for this reason, the degree of authenticity is indicated by a code.

The authentication code has three parts, separated by colons. The first part indicates the "phonological and orthographic authentication". In the example in Figure 73, the number 8 indicates a "word whose development is obscure, and whose spelling is derived wholly or partially on textual evidence" (GKK:

13). The second part indicates the items "attestation". In the example in Figure 73, "*MP*" indicates that this item is found in Middle Cornish and in Cornish toponyms. The third part of the authentication code indicates the item's frequency of occurrence in the corpus. In the example in Figure 73, the number "2" indicates that this item occurs between 2 and 3 times and "(*Origo Mundi 779,788*)" indicates that it is found in the text *Origo Mundi* at lines 779 and 788. The authentication code does not serve the purpose of distinguishing the lexeme from its homographs and might better be situated in the comment of the entry where other etymological information is found.

The pronunciation field is only included for a small number of the entries because George's Kernewek Kemmyn orthography, that is used for the head words, is intended to be phonemic. The pronunciation field is used for words with irregular stress and for words which are spelled similarly to their English translation equivalents but are pronounced differently. It can be seen from the example in Figure 73 that the transcription is narrow phonetic rather than the broad phonemic transcription more commonly found in dictionaries.

Figure 74 shows the part-of-speech markers found in George's *Gerlyver Kernewek Kemmyn* (GKK).

AJ	adjective
aj	adjective (uncertain)
AV	adverb
CJ	conjunction
CN	collective noun
DA	definite article
DN	dual noun
FN	feminine noun
fn	feminine noun (uncertain gender)
HN	noun, masculine or feminine
IJ	interjection
MN	masculine noun
mn	masculine noun (uncertain gender)
NC	number, cardinal
NO	number, ordinal
NP	personal name
PF	prefix
PH	phrase
PL	plural
PN	pronoun
PP	preposition
PV	part of verb
SF	suffix
VN	verbal noun
VP	verbal particle

Figure 74 Part-of-speech markers in GKK

Of the entries in the *Gerlyver Kernewek Kemmyn* (GKK), 301 are given no part-of-speech field. It is not clear why there is no part-of-speech attribution for these entries. It may simply be due to carelessness on the part of the compiler.

Some of the part-of-speech attributions that George (GKK) gives are questionable. For example, he marks **avel** (‘like’, ‘as’) as *AV* (adverb). However, **avel** inflects as a preposition and not as an adverb and would, therefore, be better classified as a preposition. George marks **heb** (‘without’, ‘lacking’) as *CJ* (conjunction). However, on morphological grounds, **heb** would be better classified as a preposition, since, like **avel**, it is also inflected

as a preposition.

George (GKK) does not reserve *NP* (personal name) for patronyms only; some toponyms such as **Bosveneghi** (‘Bodmin’) are also marked *NP*. Not all personal names, however, are marked *NP* by George; **Mongvras** (name of a devil), for example, is marked *FN* (feminine noun), and **Mighal** (‘Michael’) is marked *MN* (masculine noun). Toponyms, furthermore, are frequently marked by George as *MN* (e.g. **Chanel**: ‘The English Channel’) or *FN* (e.g. **Breten**: ‘Britain’). George marks **Iseldiryow** (‘Netherlands’) *PL*, but omits to mention whether it is a masculine or feminine plural. George is also inconsistent in the way in which he attributes part-of-speech to hydronyms; he marks **Tamer** (‘River Tamar’) as *mn*, but **Fowi** (‘River Fowey’) as *NP*. Of course a proper noun may be either masculine or feminine and all onomastic nouns in fact need to be marked as either masculine or feminine in the dictionary.

Unlike Williams (LCB) and Morton Nance (NCED), who distinguish verbs and their homographic derived nouns as separate lexemes, George (GKK) prefers to mark verbs, *VN* (i.e. verbal noun). However he also marks some verbs *MN* (masculine noun) or *FN* (feminine noun) as well. For example the single entry for **skila** is described as “*FN* reason, cause *VN* be the cause of”. Since it is necessary to state whether the verb, when it is acting as a noun, is masculine or feminine, one needs to do this for all verbs, which George fails to do. Of course, the distinction between the verbal and nominal use of the verbal noun also entails a semantic difference which requires a different set of translation equivalents.

George (GKK) marks some nouns *CN* (collective noun), for example, **teythi** (‘attributes’, ‘faculties’). However he again omits to mention whether these collective nouns are masculine or feminine. Some nouns are marked by George as *PL* (plural), for example, **ympynnyon** (‘brains’); again he does not mention the gender.

George (GKK) is inconsistent in his choice of base form for the noun. In the case of nouns that have both collective and singulative forms, sometimes he chooses the collective as the base form and other times he chooses the singulative as the base form. Thus we find the head word **ros** (‘roses’) marked *CN* (collective noun) followed by +**enn** to show its singulative ending. On the other hand, the head word **bodhenn** (‘a corn-marigold’) appears with its singulative ending marked *FN* (feminine noun). Its collective form, **bodh**, is not given anywhere in the *Gerlyver Kernewek Kemmyn* (GKK). George (GKK) gives separate entries for **brialli** (‘primroses’), which he marks *CN* (collective noun) and its singulative form, **briallenn** (‘a primrose’), which he marks *FN* (feminine noun).

Occasionally George (GKK) gives separate main entries to a singular noun and its plural. For example, there are main entries for the singular noun, **Kristyon** (‘Christian’), and also for its plural **Kristonyon** (‘Christians’). These two entries are adjacent in the word list. There seems to be no good reason why, in this case, the plural form should not, therefore, be given within the entry for the singular form.

In Gendall‘s *A Practical Dictionary of Modern Cornish* (PDMC) of 1997, the

head word list consists of base forms and irregular oblique forms. Appendices are provided with a chart of mutations and the paradigms of verbs, pronouns and prepositions. The lemma may include fields for the head word, a homograph number, pronunciation, part-of-speech and oblique forms.

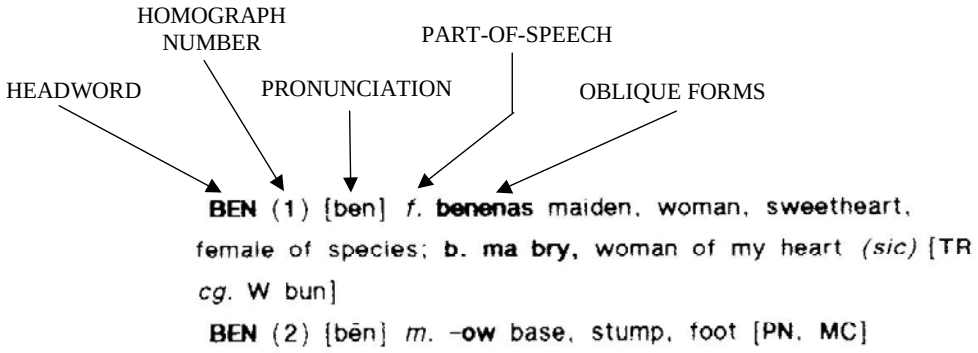


Figure 75 The lemma in PDMC

Variant spellings of the base form are cross-referenced to the preferred canonical form. Words which cause initial mutation in the following word, are marked with an asterisk.

In George's *New Standard Cornish Dictionary* (NSCD) of 1998, the head word list consists of base forms, irregular oblique forms, inflected forms of lexemes which are not found in their base form, suffixes, prefixes and stems. No appendices containing the paradigms are included. The lemma is simpler than in George's GKK and may include fields for the head word, a disambiguator, part-of-speech, and oblique forms (Figure 76). The disambiguator field still serves no useful purpose.

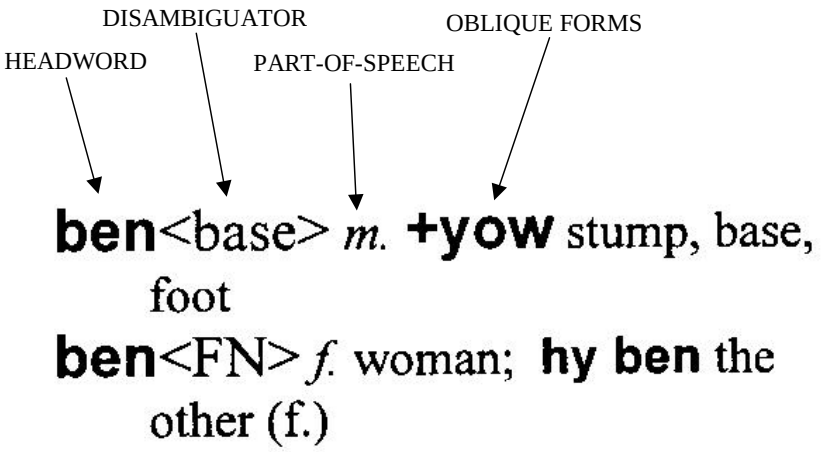


Figure 76 The lemma in NSCD

5 Methodology of Corpus Lemmatisation

The traditional method of organising dictionary corpus material is to record data on slips of card and file these by lemma. Zampolli (1981: 242) recommends that manual lemmatisation is best carried out by working on concordances of the graphic word forms. The introduction of electronic processing, however, facilitates the process of lemmatisation. Martin, Al and van Sterkenburg (1983: 81) describe software to assist in the production of lemmatised indices, concordances and frequency lists. They maintain that such programs generally consist of three components; a lexicon, a rule component and a filter. The lexicon consists of free morphemes and affixes. The rule component governs the combinatorial associations possible. The filter reduces the number of phenomena managed by the lexicon and the rule component.

The lemmatisation of Welsh shares much in common with Cornish. Cornish and Welsh share similar morphological systems and both undergo initial consonant mutation. Morris Jones (1983-1984) describes a method for the automatic tagging and lemmatisation of a corpus of Welsh child language. This method employs a computerised dictionary that relates word-types to their base forms. Morris Jones (1983-1984) notes two problems that arise from the use of a computerised dictionary for corpus lemmatisation. The first concerns elision and/or assimilation that results in contraction so that two lexemes are fused in a single word type. It was found necessary to separate the lexemes in these contractions. The second problem concerns the disambiguation of homonyms. Disambiguation was implemented by the

computer by examination of contexts.

Sondrup (Sondrup & Inglis 1982) combined three methods in preparing the lemmatised concordance of *Faust I*. First a concordance of graphic forms was used as a basis for disambiguation of homographs. Homographs were then tagged with codes in the text and a lemmatisation dictionary was constructed. Finally the concordancer was run on the tagged text with reference to the lemmatisation dictionary.

Eyes & Leech (1992: 126-7) describe the Lancaster Database of Text Corpora, one of the aims of which is to create a semantically analysed corpus and software for semi-automatic semantic tagging. Automatic grammatical tagging is thus enhanced by identifying lemmata. This is accomplished with the aid of databases of information about the contextual behaviour of lexical items compiled from large text resources. Knowledge thus obtained about the idiomaticity of language then aids sense resolution and permits the encoding of primary semantic features.

5.1 Lexeme tagging

Most corpus tagging has been concerned with syntax. Francis (1980: 198) describes how the words in the BROWN Corpus were tagged according to their syntactic function. This separates most homographs of English, though not those of identical word class. True lemmatisation of a text corpus involves inserting a tag to identify each lexeme. The system of lexeme-tags must provide a single unique code for each lexeme. An entire paradigm is thus

unified.

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. English, for example, has many compounds written as separate words but also permits nonce-compounds to be hyphenated. Lexical forms which are separated in current spelling (*i.e.* multi-word lexemes) may be treated as units and orthographically united forms, such as enclitics, may be decomposed (Zampolli 1981: 242 *ff.*).

In the SUSANNE Corpus (Sampson 1993), each token of the original text is placed on a new line, terminating in a new line character. Each line has six fields separated by tabs; reference, status, word tag, token, lemma and parse (see Figure 77).

```
A07:0050i      -       AT      the     the     [Ns[G[Nns.
A07:0050j      -       NNS1c   Mayor   mayor   .Nns]
A07:0050k      -       GG      +<apos>s        -       .G]
A07:0060a      -       VVNv    reported        report  [Tn[Vn.Vn]Tn]
A07:0060b      -       NN1c    plan    plan    .Ns]P]Ns:s]
```

Figure 77 Extract 1 from SUSANNE corpus

Tokens in the SUSANNE Corpus are often smaller than graphic words. For example punctuation marks and the apostrophe-s suffix are treated as separate words and assigned lines of their own. Some graphic words in the original text have been split in SUSANNE. The + symbol occurs as the first byte of the word field to show that the item was not separated in the original text from the immediately-preceding text segment by white space. This provides a means of representing clitics such as the genitive *s* in the example above.

Semantic annotation of the SUSANNE Corpus is undertaken at three levels, the lemma field, the marking of grammatical idioms and function tags in the parse field. The orthographic forms used in the lemma field, are based on the head words found in the 3rd edition of the *Oxford Advanced Learner's Dictionary of Current English* (OALD3). A hyphen is placed in the lemma field to represent numerals and punctuation marks. Grammatical idioms are dealt with in the parse field. In the following extract, "in touch with" is treated as a grammatical idiom equivalent to a preposition, for which the word tag is II. The nonterminal node dominating the sequence has a form tag consisting of an equals sign suffixed to the corresponding word tag. The individual words composing the grammatical idiom are not word tagged in their own right, but receive tags with numerical suffixes reflecting their membership of an idiom. The sequence "in touch with" is form tagged II=, and the words "in", "touch", and "with" in this context are word tagged II31 II32 II33. (see Figure 78)

```
A07:0250k      -       PPHS1m  he      he      [Nas:s.Nas:s]
A07:0250m      -       VVDv    made    make    [Vd.Vd]
A07:0250n      -       ATn     no      no      [Ns:o.
A07:0250p      -       NN1c    attempt attempt .
A07:0250q      -       TO      to      to      [Ti[Vi.
A07:0250r      -       VV0v    get     get     .Vi]
A07:0260a      -       II31    in      in      [P:e[II=.
A07:0260b      -       II32    touch   touch   .
A07:0260c      -       II33    with    with    .II=]
A07:0260d      -       NP1m    Carmine Carmine [Nns.
A07:0260e      A       NP1i    G.      -       .
A07:0260f      -       NP1s    De      De      .
A07:0260g      -       NP1s    Sapio   Sapio   .
A07:0260h      -       YC      +,      -       .
A07:0260i      -       AT      the     the     [Ns@.
A07:0260j      -       NP1g    Manhattan       Manhattan       [Nns.Nns]
A07:0260k      -       NN1c    leader  leader  .Ns@]Nns]P:e]Ti]Ns:o]S]
A07:0260m      -       YF      +.      -       .
```

Figure 78 Extract 2 from SUSANNE corpus

The parse field also includes function tags. These identify roles such as surface subject, logical object and time adjunct. In the extract above "he" is marked **:s** to indicate that it is the logical object. The phrase, "no

attempt to get in touch with Carmine G. De Sapio, the Manhattan Leader", is marked **:o**, to indicate that it is the logical direct object. And the phrase "in touch with Carmine G. De Sapio, the Manhattan Leader" is marked **:e** to indicate that it is the predicate complement of the subject.

5.2 Lemmatisation databases

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. 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. Morphological segmentation involves the automatic matching of raw text against a dictionary of morphemes (Zampolli 1981: 242 *ff.*). Hellberg (1972: 209), however, notes that existing dictionary lemma lists cannot cope with newspaper text corpora since these contain many neologisms.

Jones and Sondrup (1989: 495) suggest a method for the construction of a look-up dictionary. Initially two word lists are generated from the corpus, each containing identical word forms and a brief context for each item. Then each entry is checked and the second list is converted to represent the lemmata corresponding to the first list.

Bien (1981) proposes a method by which the canonical form can be indexed to its grammatical words by means of a relational database written in Prolog,

a logical computer programming language. The canonical form, which he calls a ‘morphological word’, and the grammatical word, which he calls a ‘graphemic word’, are defined as entities with certain properties. Indicator names represent binary relations holding between entity identifiers and the appropriate indicator values. The dictionary, then, consists of relationships represented in the form of simple kinds of logical formulae, stored directly in the computer. One benefit of this approach is that incomplete information can be represented. The values which are known are entered into the database and later new values can be added as they are discovered by the researcher.

It is worth examining two examples in particular, because they share similar demands and difficulties with the Corpus of Cornish. Marinone (1981) describes the preparation of a concordance to Latin Grammarians from the 2nd to the 9th centuries. The corpus consists of works by different authors with variations in spelling and forms. A means was sought to account for the evolution of the language, to organise the data to allow flexible access and provide an unlimited capacity for adding fresh data. The resulting system incorporates the following features:

1) a high number of lemmata;

2) relationships between lemmata and lemma forms; in other words, the presence of indicators which make it possible to move from the lemmata and the information contained in them to the forms;

3) relationships between lemma forms and lemmata; the presence of indicators which make it possible to move from the forms and

the information contained in them to the lemmata;

4) all the spelling and inflexional variants attested for the lemmata and forms;

5) direct access to the entire collection of data formed by the sum total of automatically processed texts and retrieval of selected information;

6) capacity to automatically increase the number of lemmata by adding all new occurrences to the lexicon whenever a new text is processed.

The system provides a tool for lemmatisation of texts as well as providing access to heterogeneous data and establishing relationships between those data.

Busharia (1979: 133*ff.*) describes the process of computerized lemmatisation of non-vocalised Hebrew texts for the *Historical Dictionary of the Hebrew Language* (HDHL). There are a vast number of forms for a single lexical item, which do not appear in one alphabetically consecutive group. Lexical particles and pronominal elements may be affixed to other items. Orthography is not at all uniform, in other words a single item may be spelt many different ways. Finally, there are a large number of homographs, this presents the main problem for computer-assisted lemmatisation. These features are all present in the Corpus of Cornish.

Busharia (1979: 136) concludes that no automatic system can replace the lexicographer's responsibility for double-checking the computer's output. Since unknown forms and lexical items may be expected, no computer

software can possess all possible language forms. And lastly, the best system makes minimal demands on the lexicographer and maximum exploitation of the computer. Furthermore, he rejects the mechanicographization of morphology, since by this method a number of lemmata are offered for each word, which results in more work for the lexicographer.

The system described by Busharia (1979: 136-8) involves the compilation of a bank of graphical forms indexed to their corresponding lemmata. 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.

Two factors help to increase the efficiency of the system. Firstly, lemmata in the bank are given a rating according to their frequency of occurrence. This is regularly updated with each new text that is processed. The most frequent lemmata are suggested first by the computer. Secondly, it is observed that different historical periods reflect different orthographies. So separate form banks for different historical periods are created. This last factor is also applicable to the case of Cornish, where in particular there is a distinction between Middle and Modern Cornish orthography.

5.3 VOLTA: a method developed for the Corpus of Cornish

The Corpus of Cornish is diachronic. It contains considerable orthographic variation and segmentation is inconsistent. Existing dictionaries of Cornish suggest that approximately 9,000 lemmata can be obtained from the

corpus. In order to make any kind of study of the Cornish lexicon, it is first of all necessary to determine the inventory of lexemes that are attested and the various graphic forms that may be united under those lexemes.

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. Furthermore 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 (NCED) and George (GKK) 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 (NCED) gives lemma status to MAGATA (‘also’), writing it as one word, unhyphenated, **magata**. George (GKK) does not give an entry for MAGATA, but gives separate entries for **maga**<as> and **da**<good>. The meaning of MAGATA is , however, not transparent from the combination of the meanings of the two items of which it is comprised. MAGATA thus requires its own entry.

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. Nor is automatic lemmatisation based on morphological rules feasible because 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.

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.

Lemma lists used by modern Cornish lexicographers (NCED; GKK) provide an initial guide to segmentation. The lemma list from George's GKK was adapted to provide a set of lexeme tags. Additional lexeme tags had to be created for lexemes that are not listed in the GKK. It is, however, important that these lexeme tags are distinguished from dictionary lemmata. The lexeme tags employed in the Corpus of Cornish, serve purely to identify lexemes. Unlike lemmata, they are not intended to represent a base or canonical form of the lexeme. Furthermore a dictionary may employ more than one lemma per lexeme to allow for orthographic variation and/or irregular forms. Thus the

lemma serves to identify the lexeme within the structure of the dictionary. There is one lemma per dictionary entry and since the same lexeme may be represented by more than one entry in the dictionary, a single lexeme may have multiple lemmata. For example George (GKK) gives separate entries for **toll** and its irregular plural **tell**. Such instances are, however, united under a single lexeme tag in the tagged corpus.

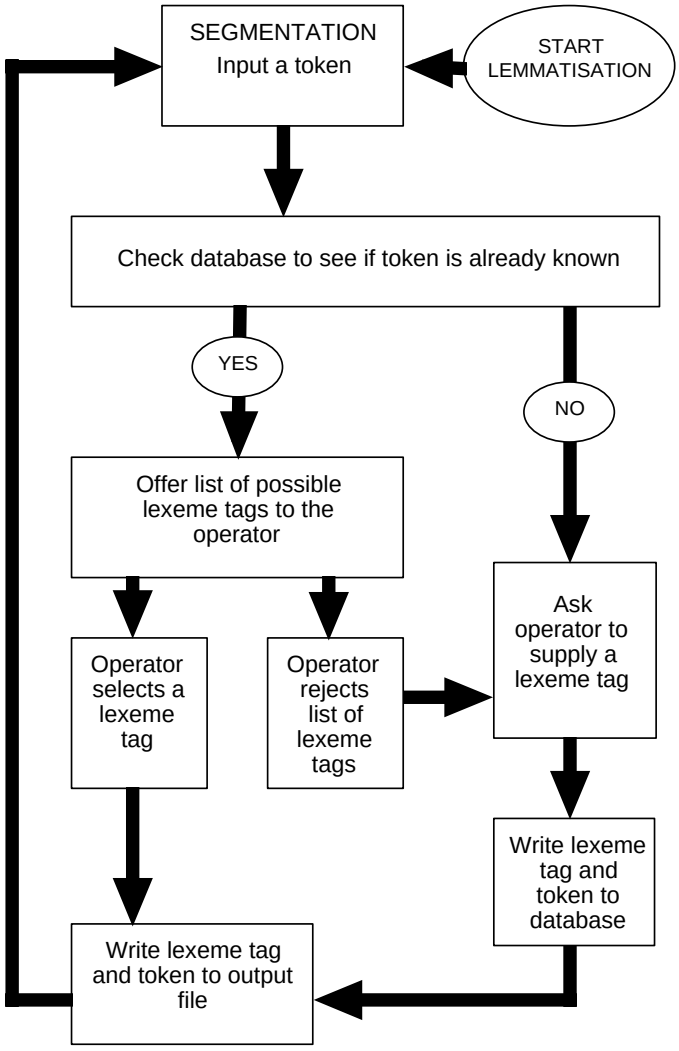


Figure 79 *VOLTA* algorithm

Computer assisted tagging of the corpus was achieved with the aid of a

specially written program called *VOLTA* (Vertical Output Lemma Tagging Aid). Figure 79 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 80 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.

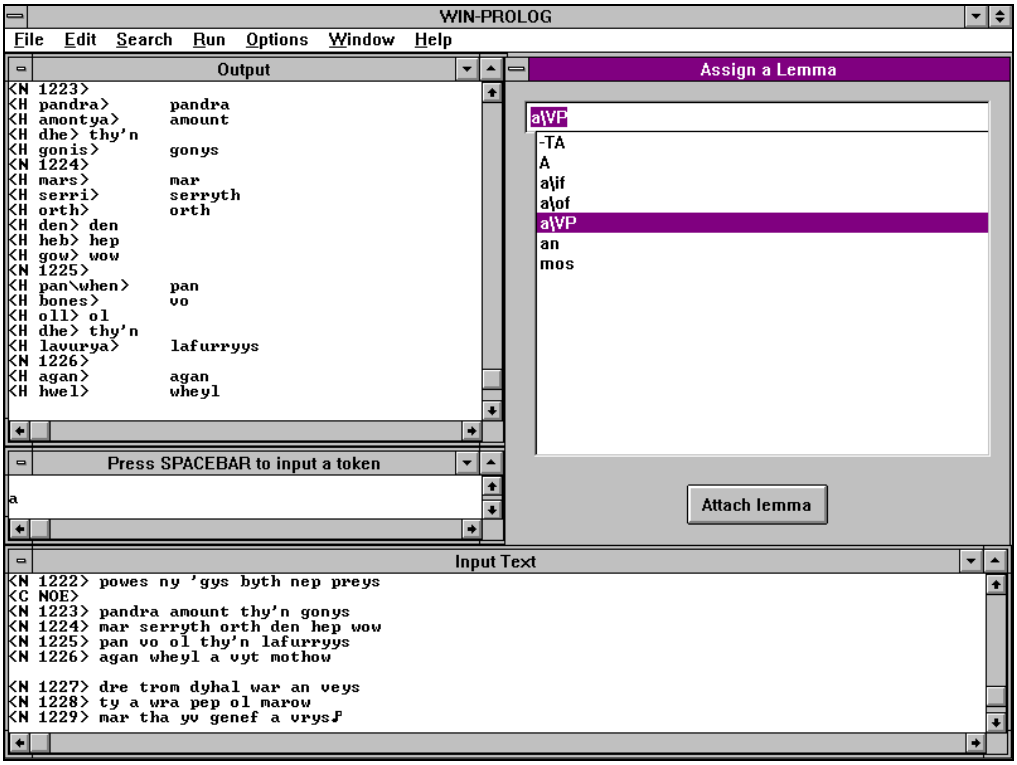


Figure 80 *VOLTA* screen during lemmatisation process

The database consists of the lexeme tag and its type entered in the form, l(*Lexeme-tag*, *Type*). Below is the database generated from William Allen's "Cornish Rhyme" (*Tonkin Manuscripts B*: 207c) a short text of 23 running words in length (see Figure 81).

```
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 81 *VOLTA* lemmatisation database

The lemmatised output file is arranged vertically with one word token per line. Comments are placed in square brackets. 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 (see Figure 82).

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.

```
[William Allen’s “Cornish Rhyme” (Tonkin Manuscripts B: 207c)]
<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 82 *VOLTA* lemmatised output

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 83 shows such a dictionary generated from William Allen’s “Cornish Rhyme” (*Tonkin Manuscripts B*: 207c), 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 83 *VOLTA* dictionary of base and oblique forms

Finally contexts may be accessed via *Micro-OCP* concordancing software. Below is a complete lemmatised KWIC concordance to William Allen’s “Cornish Rhyme” (*Tonkin Manuscripts B*: 207c) (see Figure 84).

```
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 84 Lemmatised KWIC concordance

5.4 Normalisation

Inconsistent orthography is a common problem for the corpus linguist especially in relation to older texts or those that represent diverging dialectal varieties. Rissanen (1994: 75) complains that the enormous richness of variant spellings in the Helsinki Corpus causes problems for the study of syntax or lexis. Markus (1994: 46) maintains that when compiling a Middle English prose corpus, the multiplicity of spelling variants is the main point to be considered. He reports a total of 31 variant spellings of WHEREFORE in a single random homily text of ICAMET (Innsbruck Computer

Archive of Middle English Texts). In relation to the English Century of Prose Corpus, Milić (1994: 66) points out that in the early days of printing, spelling was the prerogative of the printer which resulted in anything but uniformity. Within the corpus of Cornish, orthographic variation is a major consideration. For example, we find fifteen orthographic variants of the Cornish word for 'flesh': *chîc*, *cîg*, *cyc*, *gîc*, *gyc*, *gyke*, *kig*, *kìg*, *kîg*, *kyc*, *kych*, *kyek*, *kyg*, *kyk*, *kyke*.

A number of writers have reported the value of working with normalised texts in relation to diachronic corpora (Hickey 1994; Markus 1994; Milić 1994; Rissanen 1994). A normalised text is one in which orthographic variants of grammatical forms are replaced by single forms by external consensus. These single forms may be later standardised forms or alternatively may be arrived at by decision of the corpus compilers (Hickey 1994: 169). The advantages of normalised text include readability; the text may be approached without too much linguistic difficulty and is scholarly transparent. Furthermore normalised text improves access and is more user-friendly.

Medieval scholars sometimes object to normalisation of historical documents. It is by no means clear which norm to take with regard to the process of normalisation. Rissanen (1994: 75) considers that the "normalisation of Old and Middle English writings would necessitate a large number of awkward compromises and would, in all probability, produce a strange-looking hybrid text." However a normalised text need not be substituted for the original text but instead appended to it (Hickey 1994: 169; Markus 1994: 48).

Hickey (1994: 169 *ff.*) describes an algorithm for normalising a text. First a database is prepared which relates orthographic variants to their normalised forms. A program then compares the text word by word with the database and replaces each token with its normalised form. There are two problems with this approach to normalising the orthography of a text. Firstly, Hickey's algorithm, does not take account of homographs; in other words, a single word type in the original orthography may correspond to more than one normalised form in the database. For example, such a database compiled from Jordan's *Gwreans an Bys* includes 293 homographs of this kind. Thus the word type "the" in *Gwreans an Bys* potentially corresponds to 7 different normalised word types: *dh'y, dhe, dhe', dheu, dhy, dhy'* and *thy'*. Secondly, there is a bootstrapping problem; in other words, in order to construct such a database, one has first to put the corpus into normalised orthography.

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. They are difficult to read in this form. In order to prepare the corpus for analysis by computer, it is necessary that it be keyed in using the modern orthographic conventions of a computer keyboard. Another point to be considered is that the original spelling of the texts is not consistent, even normally within a single text. 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.

Some of the Cornish texts have been published in normalised orthography. There are three normalised spelling systems currently used by Cornish language revivalists: Unified Cornish (Kernewek Unys), Common Cornish (Kernewek Kemmyn) and Modern Cornish (Cornoack Nowydga). Unified spelling was evolved by Morton Nance for modern students of Cornish and is embodied in his *Cornish for All* (1929) and his dictionaries (ECD2, NCED, ECD3, CED). Common Cornish (Kernewek Kemmyn) is a more recent orthography devised by Ken George (1986).

In order to compare the efficiency of lemmatisation of a normalised text with a text in its original orthography, we will examine the lemmatisation of Jordan's *Gwreans an Bys*. In its original orthography, *Gwreans an Bys* contains 3,310 word types. In normalised orthography (Kernewek Kemmyn), *Gwreans an Bys* contains 2,218 word types. As a result, a lexical database that indexes all the forms of a lexeme under its lemma is smaller for normalised orthography than for the original orthography of *Gwreans an Bys*. In fact the lemmatisation database for *Gwreans an Bys* in normalised orthography contains 2,217 entries as compared with 3,309 entries in the equivalent database in original orthography.

There is also greater incidence of homography in the original orthography of *Gwreans an Bys* compared with its normalised version. Figure 85 shows the incidence of homography in original and normalised versions of *Gwreans an Bys*. Thus we see that in original orthography, there 160 word types that could be classified under two possible lemmata. In the normalised version, by

comparison, there 123 word types that could be classified under two possible lemmata. As a result, there are 196 word types that require disambiguation as compared with 149 in normalised orthography.

Number of homographs	***Number of Types***	
	Original Orthography	***Normalised Orthography***
10	*1*	*0*
9	*1*	*1*
8	*0*	*1*
7	*0*	*0*
6	*3*	*1*
5	*0*	*3*
4	*9*	*3*
3	*22*	*17*
2	*160*	*123*
Total requiring disambiguation	***196***	***149***

Figure 85 Incidence of homography in original and normalised versions of *Gwreans an Bys*

5.5 Lemmatisation rules

For normalised texts, morphological rules may be invoked to develop computer algorithms that achieve partial lemmatisation. Garside, Leech & Sampson (1991: 152-55) describe the automatic lemmatisation of the LOB Corpus, by means of a program that assigns inflexional or morpho-syntactic variants to lexical classes and sums together the frequencies of the member word-forms of each morpho-syntactic paradigm. The methodology involves taking advantage of regularities in inflexional paradigms to implement an affix stripping procedure (*cf.* Hellberg 1972).

Mills (1992: 31 *ff.*) describes how lemmatised concordances can be produced from Cornish texts which have been normalised using Morton Nance's

Unified Cornish. Lemmatisation takes places in two stages. First a word list of all the types in the text is generated. This word list is then searched for all the variant forms that the chosen lexeme assumes in the corpus. Mutated and inflected forms of the chosen lexeme are identified by reference to Morton Nance's NCED and Brown's (1984) grammar. The variants that have been identified are then entered into the search request in the concordancer and a concordance produced. In the second stage, homographs are manually separated out of the concordance. Rules governing mutation of initial consonants are helpful in disambiguating homographs.

Although inflectional variants of the same lemma can be automatically grouped together by computer, this only succeeds with items that form regular paradigms (Francis 1980: 208). Lemmatisation systems, therefore, frequently combine algorithms with data tables containing black-list entries in order to take care of word forms that cannot be lemmatised by affix stripping (Knowles 1983: 185; Garside, Leech & Sampson 1991: 152-55).

The graphic word does not distinguish between homographs, a fact that detracts from the value of frequency tables. One solution is to separate homographs before automatic lemmatisation with the aid of a KWIC concordance (Hellberg 1972). Alternatively homographs may be automatically resolved, according to their part of speech, by applying rules governing the immediate context of the homograph. Numbers may then be used to differentiate the homographs (Zampolli 1981: 242 *ff.*).

It has been frequently observed that automatic lemmatisation is not normally

100% accurate (Hellberg 1972: 212; Knowles 1983: 185; Martin, Al and van Sterkenburg 1983: 81). It therefore has to be followed by manual checking.

The direct application of a computerised morphological rule base to the historical corpus of Cornish is not practicable. There are two reasons for this. Firstly, there is a bootstrapping problem. No description currently exists of the morphology of historical Cornish, in its original orthography, that would be adequate for the purposes of a lemmatisation database. In order to compile such a database, one would, in fact, need to first lemmatise the corpus. Secondly, due to the highly capricious spelling practices found throughout the corpus, the application of a morphological rule base for the purposes of base form lemmatisation proves to be very unreliable.

The application of a computerised morphological rule base for lemmatisation becomes more feasible if the corpus is converted to normalised spelling. Lemmatisation of the corpus in its original orthography may be lemmatised by first aligning the corpus with its normalised version, and then applying the morphological rule base to the normalised version.

There are two algorithms for applying morphological rules to the corpus in order to achieve lemmatisation. The first algorithm (Figure 86) involves applying morphological rules to each word token in the corpus to determine its base form. Unless the morphology of the language in question is extremely simple, this algorithm is inefficient since a single word type may occur many times throughout the corpus and each time a word type is encountered, it is processed through the entire database of morphological rules.

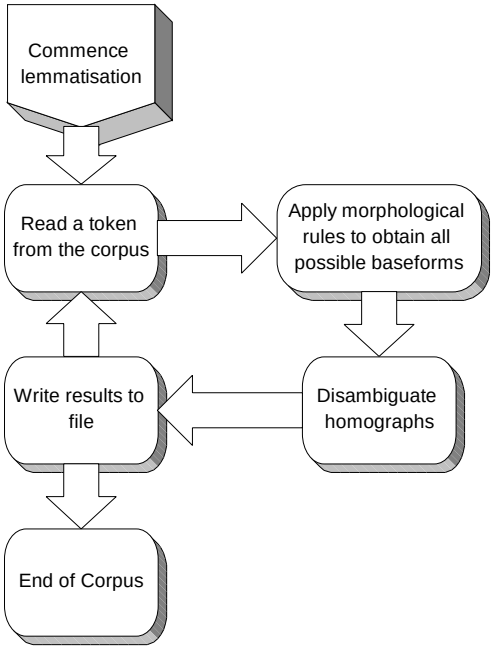


Figure 86 Morphological lemmatisation algorithm 1

The second algorithm (Figure 87) involves applying morphological rules to each word type in the corpus in order to determine its base form and thereby construct a look-up dictionary. In the case of Cornish, with its fairly complex morphology, this is the more efficient algorithm. One could manually create such a lemmatisation database of all the items contained in the *Gerlyver Kernewek Kemmyn* (GKK). However this would not be an efficient use of human resources, since entries would be created for items that are not actually attested in the corpus. It is more efficient in terms of human resources to first use the computer to generate a list of all the word types that are attested in the corpus. Then a computer program applies morphological rules to generate a database of base forms for each word type.

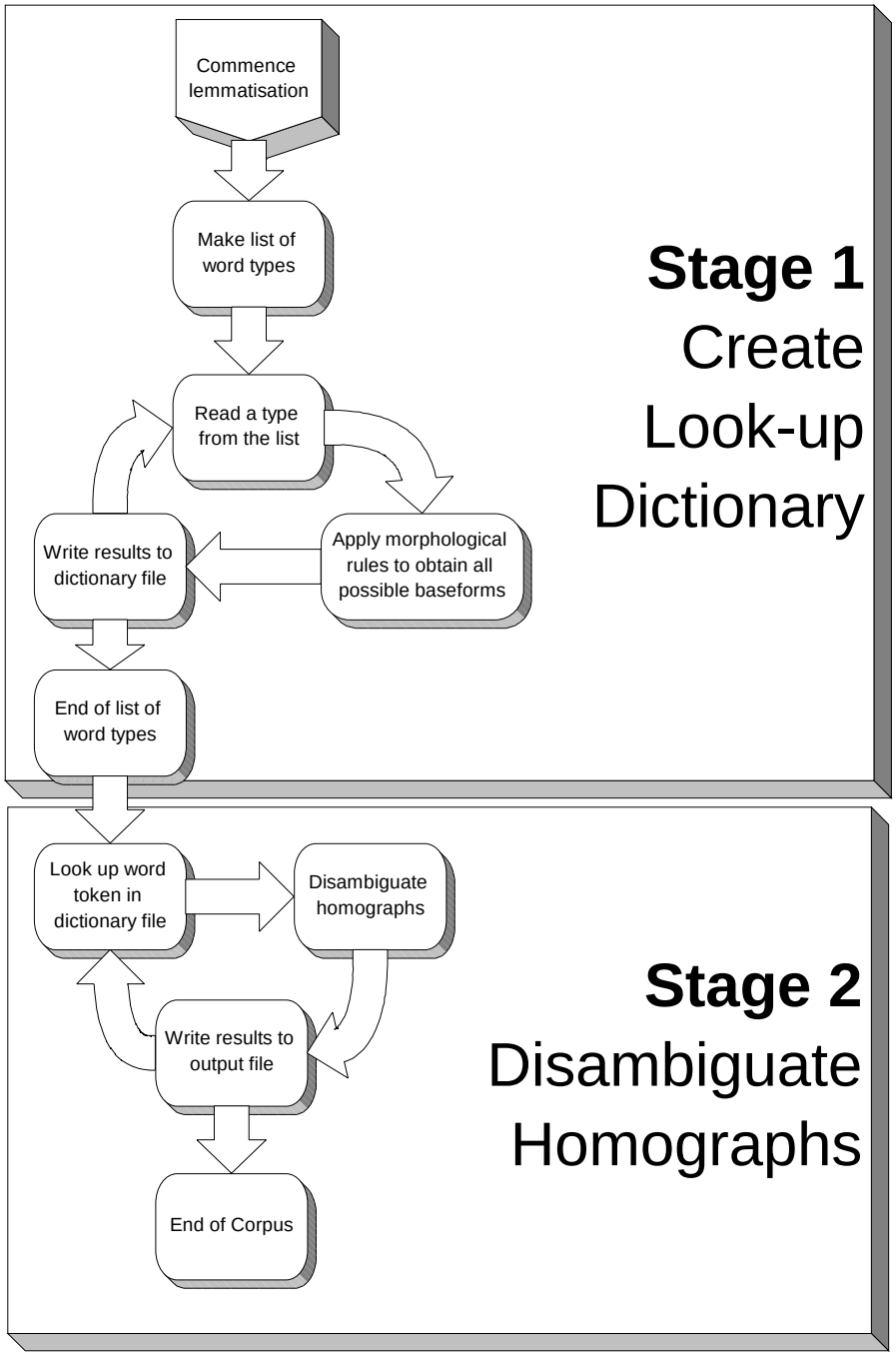


Figure 87 Morphological lemmatisation algorithm 2

The second of these algorithms was tested on a corpus in normalised spelling (Kernewek Kemmyn). The corpus includes the following texts:

1. *The Charter Endorsement,*

2. *Passhyon agan Arloedh,*
3. *The Ordinalia,*
4. *Bewnans Meryasek,*
5. *The Tregear Homilies,*
6. *Gwryans an Bys.*

In total, this corpus is comprised of 123,163 word tokens. The corpus yields a word type list of 9,611 types. The computer was used to search the word type list for items that ended with nominal plural suffixes. It was also necessary to take into account that word types may include initial consonant mutation. So by a combination of nominal plural suffix stripping and conversion of initial consonants to take into account possible mutations, a list of 3,853 word types paired with their candidate noun base forms was generated. On examination, of these 3,853 candidate noun base forms, only 326 (i.e. 8%) are actually noun base forms. The inefficiency of this algorithm is due to the fact that a word type may coincidentally terminate in a string that shares the same form as a nominal plural suffix. Thus this algorithm removes *-s* from *ambos* ('a promise') to generate *ambo* as its noun base form. *Ambo* is not a word in Cornish at all; *ambos* is already a base form, and hence the *-s* which terminates *ambos* is not in fact a suffix. Conversion of initial consonants to take into account possible mutation also creates idiosyncratic results. This algorithm thus generates the ghost word, *gambo*, as a noun base form of *ambos*.

One way of improving the algorithm is to have it check the generated candidate base form against the set of dictionary head words in order to

determine whether it is a member of that set. By this means the list of word types paired with their candidate noun base forms is reduced from 3,853 to 1,978. The 326 actual noun base forms comprise 16% of the 1,978 candidate noun base forms. A problem with this algorithm is that it generates many candidate base forms that are not in fact nouns. For example, this algorithm removes *-es* from *anes* to generate the candidate base form *an*. However, *an* is a definite article, not a noun; and *anes* (‘troubled’) is in fact an adjective.

In order to improve the algorithm further, the generated candidate base form is checked against a set of lemmata, in which each lemma is comprised of a head word and its part-of-speech. By this means only nouns are selected by the algorithm. The list of word types paired with their candidate noun base forms is now reduced to 829. The 326 actual noun base forms comprise 39% of the 829 candidate noun base forms. The reason that this algorithm is so inefficient is that many of the nominal plural suffixes are homographic. In fact, of the 14 nominal plural suffixes, 9 are heteromorphemic (see Figure 88); in other words, they share the same form as other morphemes. For example, this algorithm removes *-ys* from *alhwedhys* to generate the base form, *alhwedh*. *Alhwedh* (‘a key’) is a masculine noun, but its plural is *alhwedhow*, not *alhwedhys*. *Alhwedhys* is, in fact, the past participle of the verb *alhwedha* (‘to lock’). Thus the suffix, *-ys*, is homographic.

-ow	nominal plural
-yow	nominal plural
-en	nominal plural, verbal suffix
--on	nominal plural
-yon	nominal plural
-yn	nominal plural, verbal suffix
-yer	nominal plural, derivational agency noun suffix
-y	nominal plural, verbal suffix
-s	nominal plural
-as	nominal plural, verbal suffix
-es	nominal plural, verbal suffix
-ys	nominal plural, verbal suffix, derivational abstract noun suffix
-ans	nominal plural, derivational abstract noun suffix
-eth	nominal plural, derivational abstract noun suffix

Figure 88 Nominal plural suffixes

In order for the system to know whether *alhwedhys* is in fact the plural of the masculine noun, *alhwedh*, it is necessary for the system to know to which nominal declension *alhwedh* belongs. Such a system no longer relies simply on suffix stripping; such a system is in effect a relational database of lemmata and their base forms and oblique forms.

There are two possible approaches to creating a computerised morphological analyser for the purpose of lemmatisation. The stochastic approach applies statistical techniques to the corpus in order to learn its morphology. The second approach involves manually constructing a morphological database from existing descriptions of Cornish morphology in the various published grammars. These two approaches were trialed on the Corpus of Cornish in its normalised orthography (Kernewek Kemmyn) version. This corpus consists of the following texts:

The Charter Endorsement,

Passhyon agan Arluth,

The Ordinalia,

Bewnans Meryasek,

The Tregear Homilies,

Gwryans an Bys.

The corpus consists of 120,993 word tokens and 9025 word types.

5.6 *The stochastic approach to generating morphological rules*

Linguistica is a computer program that was developed by John Goldsmith of the Department of Linguistics at the University of Chicago. It is a C++ program that functions as a Windows-based tool for corpus-based linguistics. The aim of the program is to learn the structure of words in any human language on the basis of a raw text. No human supervision is required, except for the naïve creation of the text. *Linguistica* tells you that a given language has a category of words that take a particular set of suffixes. The morphology that is produced by *Linguistica* consists of 3 things: a list of stems, a list of suffixes, and a list of signatures with their associated stems. A signature is the pattern of suffixes that a stem takes.

The data that *Linguistica* requires in order to construct a morphology is a corpus. Reasonable results are rapidly obtainable with a corpus of 5,000 tokens, but results are improved with 50,000 tokens, and much improved with 500,000 tokens. *Linguistica* uses 2 heuristics. The first heuristic uses weighted mutual information to search for basic candidate suffixes of the language. Mutual information is a statistical measure of the degree of relatedness of 2 elements based on the ratio between observed and expected results. The

second heuristic uses these basic candidate suffixes to find regular signatures. The system treats a signature as strictly regular if it contains more than one suffix, and is found on more than one stem. A suffix found in a strictly regular suffix is treated as a regular suffix. Only regular signatures composed of regular suffixes are retained by the system. Minimum description length is then employed to correct errors generated by these heuristics. Minimum description length (Rissanen 1987) is a very powerful and general approach which can be applied to any inductive learning task. It works on the principle that the simplest theory which explains the data is the best theory.

Linguistica was employed to analyse the normalised orthography (Kernewek Kemmyn) version of the Corpus of Cornish. *Linguistica* identified 4,045 stems, 105 regular suffixes and 983 signatures with regular suffixes. Figure 89 shows some of the stems with their signatures as identified by *Linguistica*.

kabla	*NULL s*
kaff	*av en o*
kaffe	*ns wgh*
kafo	*es*
kaif	*as*
kal	*a es s*
kales	*NULL sa*
kalett	*er*
kalkor	*yon*
kall	*a av en ewgh*
kamm	*NULL enn*
kammhyns	*eth*
kammonderstond	*ya*
kammworthyb	*is*
kampoell	*ys*
kan	*NULL a av ens ow s*
kanj	*on*
kann	*as*
kannas	*NULL ow*
kans	*NULL ow*

Figure 89 *Linguistica* stems and signitures

Let us examine the stems and their signatures shown in Figure 89.

Kabla is a verb (‘to blame’). *Kablas* (‘blamed’) is the 3rd person singular of the preterite tense. The stem is in fact *kabl-*.

Kaff- is the stem of the verb *kavoes* (‘to find’). This stem also has the allomorphs, *kav-*, *kev-*, *kyff-* and *kyv-*, which *Linguistica* does not group together. *Kaffav, kaffen* and *kaffo* are all part of the verbal paradigm of *kavoes. Kaffe* is not a genuine stem. *Linguistica* has split the words types, *kaffens* and *kaffewgh* in the wrong place. The suffixes are, in fact, *-ens* and *-ewgh*, and *kaffens* and *kaffewgh* are also parts of the paradigm of *kavoes*.

Kaifas is a personal name and a single morpheme. The *-as*, that terminates *kaifas*, is, thus, not a suffix.

Kal- is not in fact a stem in Cornish. The word types, *kala* (‘a straw’), *kales* (‘difficult’) and *kals* (‘a heap’) all consist of single morphemes.

Kales (‘difficult’) has the comparative form, *kalessa* (‘more difficult’). It would perhaps be a better analysis to say that the comparative suffix is *-a* and *kaless-* is a stem allomorph.

Kaletter (‘difficulty’) is derived from *kales* (‘difficult’). The derivational suffix is in fact *-ter* and *kalet-* is an allomorph of *kales*.

Kalkoryon (‘mathematicians’) has been correctly segmented by *Linguistica* into its stem, *kalkor* (‘a mathematician’) and its plural suffix, *-yon*.

Kalla, *kallav*, *kallen* and *kallewgh* are all part of the verbal paradigm of *galloes* with mutation of initial *g-* to *k-*.

Kamm (‘a step’) has the derivation, *kammen* (‘a way’), by addition of the suffix *-enn*.

Kammhynseth (‘injustice’) is a compound stem, formed from the morphemes, *kamm* (‘bent’), *hyns* (‘way’) and *-eth*.

Kammonderstondya (‘to misunderstand’) has been correctly analysed by *Linguistica* into the stem, *kammonderstond-* and its verbal noun suffix, *-ya*.

Kammworthybis (‘replied impertinently’) has been correctly analysed by *Linguistica* into the stem, *kammworthyb-* and its 3rd person singular preterite suffix, *-is*.

Kampoellys (‘mentioned’) is the past participle of *kampoella* (‘to mention’) and has been correctly analysed by *Linguistica*.

Kan is the stem of a vocable that includes the lexemes KAN (‘a song’) and KANA (‘to sing’). *Linguistica* has correctly segmented this vocable. Its suffixes *-a*, *-av*, *-ens* and *-s* are verbal suffixes, and *-ow* is a nominal plural suffix.

Kanjon (‘a wretch’) is a single morpheme and has been wrongly segmented by *Linguistica*.

Kannas (‘a messenger’) is a single morpheme and has been wrongly segmented by *Linguistica*. However, *kannasow* (‘messengers’) has been correctly segmented.

Kans (‘a hundred’) and its plural, *kansow* (‘hundreds’), have been correctly segmented by *Linguistica*.

Thus of the 20 stems with their signatures in Figure 89, only 9, or less than half, have been correctly segmented into their morphemes by *Linguistica*. *Linguistica* might be said to have made a good guess at Cornish morphology, but its output cannot be regarded as very reliable. *Linguistica* takes no account of vowel affection or morphophonemic alternation.

The output from *Linguistica* was used to create a lemmatisation database. First the *Linguistica* generated file of stems and affixes was converted to a Prolog database of stems and affixes. Figure 90 shows the form of the database.

```
stem_and_suffixes(‘kabla’, [‘’, ‘s’]).
stem_and_suffixes(‘kaff’, [‘av’, ‘en’, ‘o’]).
stem_and_suffixes(‘kaffe’, [‘ns’, ‘wgh’]).
stem_and_suffixes(‘kafo’, [‘es’]).
stem_and_suffixes(‘kaif’, [‘as’]).
stem_and_suffixes(‘kal’, [‘a’, ‘es’, ‘s’]).
stem_and_suffixes(‘kales’, [‘’, ‘sa’]).
stem_and_suffixes(‘kalett’, [‘er’]).
stem_and_suffixes(‘kalkor’, [‘yon’]).
stem_and_suffixes(‘kall’, [‘a’, ‘av’, ‘en’, ‘ewgh’]).
stem_and_suffixes(‘kamm’, [‘’, ‘enn’]).
stem_and_suffixes(‘kammhyns’, [‘eth’]).
stem_and_suffixes(‘kammonderstond’, [‘ya’]).
stem_and_suffixes(‘kammworthyb’, [‘is’]).
stem_and_suffixes(‘kampoell’, [‘ys’]).
stem_and_suffixes(‘kan’, [‘a’, ‘av’, ‘ens’, ‘ow’, ‘s’]).
stem_and_suffixes(‘kanj’, [‘on’]).
stem_and_suffixes(‘kann’, [‘as’]).
stem_and_suffixes(‘kannas’, [‘’, ‘ow’]).
stem_and_suffixes(‘kans’, [‘’, ‘ow’]).
```

Figure 90 Database of stems and their affixes

Next, for each stem and its respective affixes, all the possible combinations of stem and affix were generated. Thus, from the stem *kan-* and its affixes *-a*, *-av*, *-ens*, *-ow* and *-s*, the forms *kana*, *kanav*, *kanens*, *kanow* and *kans* are generated. Taking into account that the initial *k-* of the stem could be a mutation of *g-*, we can add the following putative forms to the set: g*ana*, g*anav*, *ganens*, g*anow* and *gans*.

Set theory then provides the means by which base forms and oblique forms are distinguished. Let *s* be a stem. *A* is the set of all affixes, *a*, such that *a* is an affix which occurs with *s*. $s \times A$ is the Cartesian product of the stem, *s*, and its

set of affixes, *A*. *H* is the set of all head words, *h*, such that *h* is a head word in the *Gerlyver Kernewek Kemmyn* (GKK). Then *B*, the intersection of ($s \times A$) and *H* is the set of possible base forms for the stem, *s*; and *O* the set of all oblique forms for the stem, *s*, is the difference between ($s \times A$) and the set of all base forms, *B* (see Figure 91).

s = a stem,

$A = \{a \mid a$ is an affix which occurs with $s\}$,

$s \times A = \{<s, a> \mid s , a$ in $A\}$,

$H = \{h \mid h$ is a head word in the GKK$\}$,

$B = ((s \times A) \cap H)$,

$O = ((s \times A) - B)$.

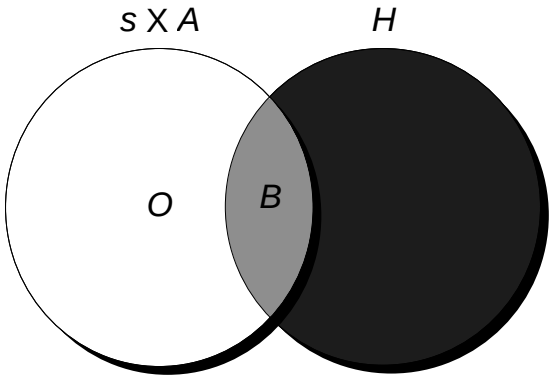


Figure 91 Venn diagram of base and oblique forms

Using these set theory operations a Prolog program was written to convert the database of stems and affixes (see Figure 90) to a Prolog database of base forms and their variant forms. The extract from this database, shown in Figure 92, shows that the base form *kabla* ('to blame') has the variant forms *kabla*

and *kablas*.

```
baseform_variants(kabla, [kabla, kablas]).
baseform_variants(kablys, [hablys]).
baseform_variants(kachya, [gach, gachya]).
baseform_variants(kala, [gal, gala, gales, galk, galow, galsa, galw, galar,
                kala, kales, kals]).
baseform_variants(kales, [gal, gala, gales, galk, galow, galsa, galw, kala,
                kales, kals, kalessa]).
baseform_variants(kaletter, [kaletter]).
baseform_variants(kals, [kala, kales, kals]).
baseform_variants(kamm, [gam, gamm, gamma, kamm, kammenn]).
baseform_variants(kamma, [gamm, gamma]).
baseform_variants(kammenn, [kamm, kammenn]).
baseform_variants(kammhynseth, [gammhynseth, kammhynseth]).
baseform_variants(kammneves, [gammneves]).
baseform_variants(kan, [ga, gal, gam, gan, gar, gara, gas, gasa, gav, gava,
                gay, gana, ganow, gans, han, hanas, haneth, hanow, hans,
                hanter, kan, kana, kanav, kanens, kanow, kans]).
baseform_variants(kana, [gan, gana, ganow, gans, kan, kana, kanav, kanens,
                kanow, kans]).
baseform_variants(kanjon, [kanjon]).
baseform_variants(kannas, [gannas, ganno, hannas, kannas, kannasow]).
baseform_variants(kans, [gan, gana, ganow, gans, gansen, gansi, ganso, gansons,
                han, hanas, haneth, hanow, hans, hanter, hansel, kan, kana,
                kanav, kanens, kanow, kans, kansow]).
```

Figure 92 Prolog database of base forms and their variant forms

Since identifying base forms is only part of the process of lemmatisation, it is also necessary to distinguish between homographic base forms. The Prolog database of base forms and their variant forms illustrated in Figure 92 was converted to a Prolog database of lemmata and their variant forms by the addition of two fields from the *Gerlyver Kernewek Kemmyn* (GKK): the distinguisher field, and the part-of- speech field. The extract from this

new database, shown in Figure 93, distinguishes three homographs of the base form, *kamm*.

```
lemma_variants(kablys, ‘’, ‘MN’, [hablys]).
lemma_variants(kachya, ‘’, ‘VN’, [gach, gachya]).
lemma_variants(kala, ‘’, ‘CN‘, [gal, gala, gales, galk, galow, galsa, galw,
              galar, kala, kales, kals]).
lemma_variants(kales, ‘’, ‘AJ’, [gal, gala, gales, galk, galow, galsa, galw,
              kala, kales, kals, kalessa]).
lemma_variants(kaletter, ‘’, ‘MN’, [kaletter]).
lemma_variants(kals, ‘’, ‘MN’, [kala, kales, kals]).
lemma_variants(kamm, bent, ‘AJ’, [gam, gamm, gamma, kamm, kammenn]).
lemma_variants(kamm, bent, ‘MN’, [gam, gamm, gamma, kamm, kammenn]).
lemma_variants(kamm, step, ‘MN’, [gam, gamm, gamma, kamm, kammenn]).
lemma_variants(kamma, ‘’, ‘VN’, [gamm, gamma]).
lemma_variants(kammenn, ‘’, ‘FN’, [kamm, kammenn]).
lemma_variants(kammhynseth, ‘’, ‘MN’, [gammhynseth, kammhynseth]).
lemma_variants(kammneves, ‘’, ‘FN’, [gammneves]).
lemma_variants(kan, ‘’, ‘FN’, [ga, gal, gam, gan, gar, gara, gas, gasa, gav,
              gava, gay, gana, ganow, gans, han, hanas, haneth, hanow, hans,
              hanter, kan, kana, kanav, kanens, kanow, kans]).
lemma_variants(kana, ‘’, ‘VN’, [gan, gana, ganow, gans, kan, kana, kanav,
              kanens, kanow, kans]).
lemma_variants(kanjon, ‘’, mn, [kanjon]).
lemma_variants(kannas, ‘’, ‘FN’, [gannas, ganno, hannas, kannas, kannasow]).
lemma_variants(kans, ‘’, ‘NC‘, [gan, gana, ganow, gans, gansen, gansi, ganso,
              gansons, han, hanas, haneth, hanow, hans, hanter, hansel, kan,
              kana, kanav, kanens, kanow, kans, kansow]).
```

Figure 93 Prolog database of lemmata and their variant forms

The database of lemmata and their variant forms illustrated in Figure 93 was applied to the list of word types found in the normalised orthography (Kernewek Kemmyn) version of the Corpus of Cornish. Figure 94 is a chart of the number of types for which a given number of lemmata are suggested.

From this chart we can see that the system suggests no lemmata for 5,142 of the word-types that are attested in the corpus. For 2,145 word-types only one lemma is suggested. The number of lemmata suggested by the system for any given word type might be as many as 36. Thus, for the word type, *fal*, the system suggests the following 36 items as possible lemmata: **bal** MN; **ball** (plague) FN; **ball** (spot) MN; **bara** MN; **bas** AJ; **bay** (bay) MN; **bay** (kiss) MN; **fall** MN; **fals** AJ; **fals** FN; **fara** MN; **fas** MN; **fashyon** MN; **fay** MN; **ma** CJ; **ma** PN; **mal** IJ; **mall** MN; **mar** CJ; **mar** mn; **mar** AV; **mars** CJ; **mas** AJ; **maw** MN; **may** CJ; **pal** FN; **pall** MN; **pals** AJ; **par** AV; **par** PP; **par** MN; **para** VN; **para** MN; **pas** (cough) MN; **pas** (pace) MN; **paw** MN.

Lemmata	***Types***	***Lemmata***	***Types***
0	*5142*	*18*	*24*
1	*2145*	*19*	*25*
2	*1005*	*20*	*25*
3	*491*	*21*	*18*
4	*224*	*22*	*14*
5	*195*	*23*	*7*
6	*153*	*24*	*7*
7	*136*	*25*	*21*
8	*103*	*26*	*7*
9	*41*	*27*	*7*
10	*35*	*28*	*8*
11	*42*	*29*	*2*
12	*40*	*30*	*1*
13	*31*	*31*	*3*
14	*39*	*33*	*1*
15	*47*	*34*	*1*
16	*21*	*35*	*1*
17	*13*	*36*	*2*

Figure 94 The number of types for which a given number of lemmata are suggested

Figure 95 shows the proportion of word types for which the system suggests 0 lemmata, 1 lemma or more than 1 lemma. It can be seen that for just

over half the word types, the system fails to suggest a lemma. For a little more than a quarter of the word types, the system suggest more than one lemma. When this occurs, human intervention is necessary to choose the correct lemma from the set offered. For slightly less than a quarter of the word types, only one lemma is suggested by the system.

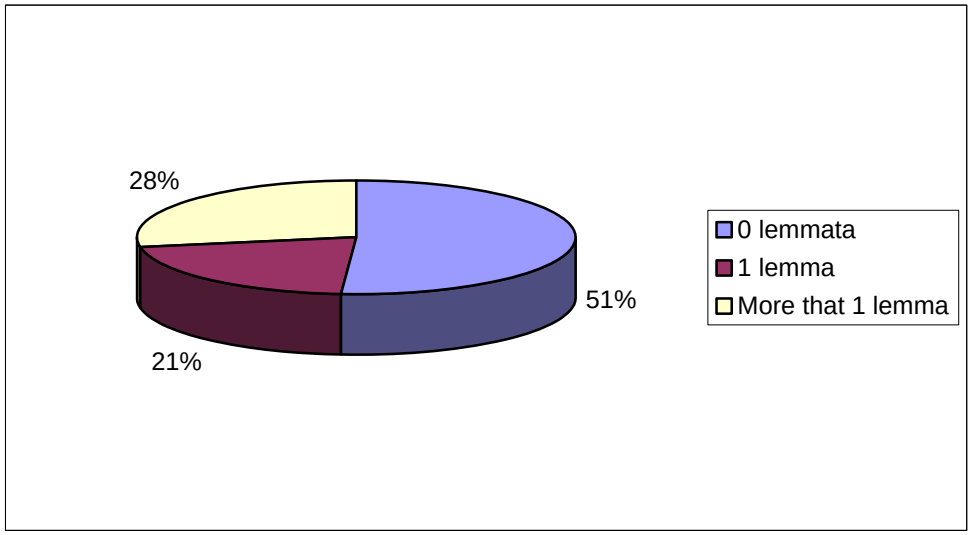


Figure 95 Proportion of word types for which the system suggests 0 lemmata, 1 lemma or more than 1 lemma

Although the *Linguistica* analysis of Cornish morphology in normalised spelling might be said to be a good guess at Cornish morphology, it has several shortcomings. Firstly it does not account for all the word types that are attested in the corpus. Secondly it is frequently ambiguous. Occasionally it is completely wrong. And finally, it does not provide any knowledge of Cornish morphology that is not already described in the literature.

Considerable human intervention in the lemmatisation process would thus be necessary if a lemmatisation database obtained from the *Linguistica*

analysis is used. The human operator would need to attribute lemmata to the 51% of word types for which the *Linguistica* analysis offers no suggestion. In the case of the 28% of word types for which the *Linguistica* analysis suggests more than one possible lemma, the human operator would need to select the correct lemma. And any suggestions for attributing lemmata to word types that the *Linguistica* analysis makes would need to be checked for accuracy by the human operator.

In the case of a language for which little or no morphological description already exists, *Linguistica* might prove to be a useful tool.

5.7 Manual creation of a morphological analyser

A morphological analyser was created manually for the purposes of head word lemmatisation. The morphological analyser was designed to be used with the Corpus of Cornish in normalised orthography (Kernewek Kemmyn), thereby overcoming the problems of capricious spelling that are found in the texts in their original form. The Corpus of Cornish in its normalised spelling version contains 9610 word types. These are analysed with regard to mutation of initial consonants and inflectional affixes. The morphological analyser consists of 8 Rules.

Rule 1 is as follows. If a word type is identical in form with the base form of a given lemma, then that word type may be classified under that lemma. This may be expressed in predicate logic as follows:

t = a word type

b = a base form

l = a lemma

L = is a lemma with a given base form

F = are identical in form

C = word type may be classified under lemma

$$\forall(t)\ \exists(b)\ F(t, b)\ \&\ \exists(l)\ L(l, b) \rightarrow C(t, l)$$

This formula is coded in Prolog as follows:

```
type_lemma_chars(Type,(Type,Dis,POS)):-
 type_freq(Type,_,_),
 lemma(Type,Dis,POS).
```

Rule 2 is as follows. If a word type is identical in form with a mutated form of the base form of a given lemma whose part-of-speech is of a type that permits mutation of the initial consonant (i.e. an adjective, adverb, noun, proper noun, number, or verb), then that word type may be classified under that lemma. This may be expressed in predicate logic as follows:

t = a word type

b = a base form

m = a mutated base form

l = a lemma

L = is a lemma with a given base form

M = is a mutation of

F = are identical in form

C = word type may be classified under lemma

$$\exists(b)\ \exists(m)\ M(m, b)\ \&\ \forall(t)\ F(t, m)\ \&\ \exists(l)\ L(l, b) \rightarrow C(t, l)$$

This formula is coded in Prolog as follows:

```
type_lemma_chars(Type,(Base,Dis,POS)):-
 permits_mutation(POS),
 type_freq(Type,_,_),
 demutate(Base,Type),
 lemma(Base,Dis,POS).
```

Rule 3 is as follows. If a word type is identical in form with the base form plus a suffix and that base form is of a given lemma whose part-of-speech permits that suffix, then that word type may be classified under that lemma. This may be expressed in predicate logic as follows:

t = a word type

b = a base form

s = a suffix

l = a lemma

p = a part-of-speech

L = is a lemma with a given base form and a given part-of-speech

P = permits

S = is the conflation of a root and a suffix

F = are identical in form

C = word type may be classified under lemma

$$\exists(b)\ \exists(l)\ \exists(p)\ L(l, b, p)\ \&\ \exists(s)\ P(p, s)\ \&\ \forall(t)\ S((b, s), t,) \rightarrow C(t, l)$$

This formula is coded in Prolog as follows:

```
type_lemma_chars(Type,(Base,Dis,POS)):-
```

```
type_freq(Type,_,_),
suffix(POS,Suffix),
append(Base,Suffix,Type),
lemma(Base,Dis,POS).
```

Rule 4 is as follows. If a word type is identical in form with a mutated form of the base form concatenated with a suffix and that base form is of a given lemma whose part-of-speech permits that suffix, then that word type may be classified under that lemma. This may be expressed in predicate logic as follows:

t = a word type

b = a base form

s = a suffix

l = a lemma

p = a part-of-speech

m = a mutated base form

L = is a lemma with a given base form and a given part-of-speech

P = permits

S = is the conflation of a root and a suffix

M = is a mutation of

F = are identical in form

C = word type may be classified under lemma

$$\exists(b)\ \exists(l)\ \exists(p)\ L(l, b, p)\ \&\ \exists(s)\ P(p, s)\ \&\ \exists(m)\ M(m, b)\ \ \&\ \forall(t)\ S((m, s), t,) \rightarrow C(t, l)$$

This formula is coded in Prolog as follows:

```
type_lemma_chars(Type,(Base,Dis,POS)):-
 (POS='MN'; POS='FN'; POS='AJ'; POS='VN'),
 type_freq(Type,_,_),
 demutate(Type,Demutated),
 suffix(POS,Suffix),
 append(Base,Suffix,Demutated),
 lemma(Base,Dis,POS).
```

Rule 5 is as follows. If a word type is identical in form with the root of a verb, whose base form includes a verbal-nominal root, concatenated with a verbal suffix, then that word type may be classified under that lemma. This may be expressed in predicate logic as follows:

t = a word type

b = a base form

r = a root

s = a verbal suffix

l = a lemma

L = is a lemma with a given base form

R = is the root of

S = is the conflation of a root and a suffix

F = are identical in form

C = word type may be classified under lemma

V = is a verb

$$\exists(b)\ \exists(l)\ L(l, b,)\ \&\ V(l)\ \&\ \exists(r)\ R(b, r)\ \&\ \forall(t)\ \exists(s)\ S((r, s), t) \rightarrow C(t, l)$$

This formula is coded in Prolog as follows:

```
type_lemma_chars(Type,(Base,Dis,'VN')):-
```

```
type_freq(Type,_,_),
suffix('VN',Suffix),
append(Stem,Suffix,Type),
lemma(Base,Dis,'VN'),
get_vb_stem(Base,Stem).
```

Rule 6 is as follows. If a word type is identical in form with a mutated form of the root of a verb, whose base form includes a verbal-nominal root, concatenated with a verbal suffix, then that word type may be classified under that lemma. This may be expressed in predicate logic as follows:

t = a word type

b = a base form

r = a root

m = a mutated root

s = a verbal suffix

l = a lemma

L = is a lemma with a given base form

R = is the root of

M = is a mutation of

S = is the conflation of a root and a suffix

F = are identical in form

C = word type may be classified under lemma

V = is a verb

$$\exists(b)\ \exists(l)\ L(l, b,)\ \&\ V(l)\ \&\ \exists(r)\ R(b, r)\ \&\ M(m, r)\ \&\ \forall(t)\ \exists(s)\ S((m, s), t) \rightarrow C(t, l)$$

This formula is coded in Prolog as follows:

```
type_lemma_chars(Type,(Base,Dis,'VN')):-
 type_freq(Type,_,_),
 demutate(Type,Demutated),
 suffix('VN',Suffix),
 append(Stem,Suffix,Demutated),
 lemma(Base,Dis,'VN'),
 get_vb_stem(Base,Stem).
```

Rule 7 is as follows. If a word type is identical in form with the root of a verb whose base form includes a verbal-nominal root concatenated with the verbal suffix, *-he*, then that word type may be classified under that lemma. This may be expressed in predicate logic as follows:

t = a word type

b = a base form

r = a root

s = verbal suffix -he

l = a lemma

L = is a lemma with a given base form

R = is the root of

S = is the conflation of a root and a suffix

F = are identical in form

C = word type may be classified under lemma

V = is a verb

$$\exists(b)\ \exists(l)\ L(l, b,)\ \&\ V(l)\ \&\ \exists(r)\ R(b, r)\ \&\ \forall(t)\ \exists(s)\ S((r, s), t) \rightarrow C(t, l)$$

This formula is coded in Prolog as follows:

```
type_lemma_chars(Type,(Base,Dis,'VN')):-
 type_freq(Type,_,_),
 append(Stem,[104|_],Type),
 append(Stem,"he",Base),
 lemma(Base,Dis,'VN').
```

Rule 8 is as follows. If a word type is identical in form with a mutated form of the root of a verb whose base form includes a verbal-nominal root concatenated with the verbal suffix, *-he*, then that word type may be classified under that lemma. This may be expressed in predicate logic as follows:

t = a word type

b = a base form

r = a root

m = a mutated root

s = verbal suffix -he

l = a lemma

L = is a lemma with a given base form

R = is the root of

M = is a mutation of

S = is the conflation of a root and a suffix

F = are identical in form

C = word type may be classified under lemma

V = is a verb

$$\exists(b)\ \exists(l)\ L(l, b,)\ \&\ V(l)\ \&\ \exists(r)\ R(b, r)\ \&\ M(m, r)\ \&\ \forall(t)\ \exists(s)\ S((m, s), t) \rightarrow C(t, l)$$

This formula is coded in Prolog as follows:

```
type_lemma_chars(Type,(Base,Dis,'VN')):-
 type_freq(Type,_,_),
 demutate(Type,Demutated),
 append(Stem,[104|_],Demutated),
 append(Stem,"he",Base),
 lemma(Base,Dis,'VN').
```

The Prolog morphological analyser runs very slowly. To analyse the 9610 word types in the normalised version of the Corpus of Cornish the program took 2½ weeks. Rules 1 and 2 are completely reliable; in other words, the output generated by these rules requires no checking. 3307 (34%) word types attested in the normalised spelling corpus are identical in form with base forms. 396 (4%) word types attested in the normalised spelling corpus are identical in form with base forms that have undergone initial mutation. Rules 3, 4, 5 and 6 are not completely reliable. The output generated by these rules, therefore, needs to be checked manually. 1383 (14%) word types attested in the normalised spelling corpus are identical in form with a base form plus a suffix. Rule 5 classifies 1672 types under 849 lemmata to provide 1901 entries (type-lemma pairs) in the lemmatisation database. After manual checking 389 of the 1901 entries were found to be erroneous. These were removed from the database to leave 1512 correct that classify 1485 types under 781 lemmata.

After checking, the morphological analyser had created a lemmatisation database containing 6112 entries that classify 4888 word types under 3624 lemmata. Figure 96 is a chart of the number of types for which a given number of lemmata are suggested. From this chart we can see that the system suggests

no lemmata for 3,489 of the word-types that are attested in the corpus. For 4015 word-types only one lemma is suggested. The maximum number of lemmata suggested by the system for one word type is 8.

Lemmata	*Types*
0	*3498*
1	*4015*
2	*642*
3	*153*
4	*49*
5	*21*
6	*5*
7	*2*
8	*1*

Figure 96 The number of types for which a given number of lemmata are suggested

Figure 97 shows the proportion of word types for which the system suggests 0 lemmata, 1 lemma or more than 1 lemma. It can be seen that for 42% of the word types, the system fails to suggest a lemma. For 10% of the word types, the system suggest more than one lemma; these will require disambiguation. For very nearly half of the word types, only one lemma is suggested by the system.

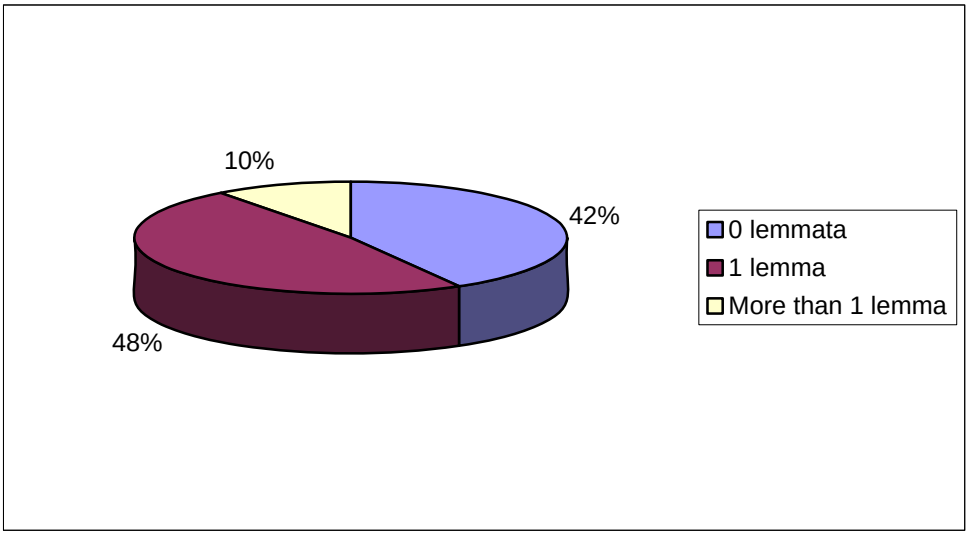


Figure 97 Proportion of word types for which the system suggests 0 lemmata, 1 lemma or more than 1 lemma

This left 3498 word types unanalysed. In order to complete the lemmatisation database, these were lemmatised manually. This manual lemmatisation was undertaken with the assistance of a concordancer, in order that each word type could be viewed in all of its contexts.

Figure 98 compares the proportion of word types for which 0 lemmata, 1 lemma or more than 1 lemma are suggested by the stochastic morphological system (using *Linguistica*) and the manually created morphological analyser. It can be seen that the manually created system outperforms the system based on *Linguistica*'s output in all aspects. The manually created system leaves fewer word types for which no lemma is suggested and far fewer types that require disambiguation as a result of the system suggesting more than one lemma.

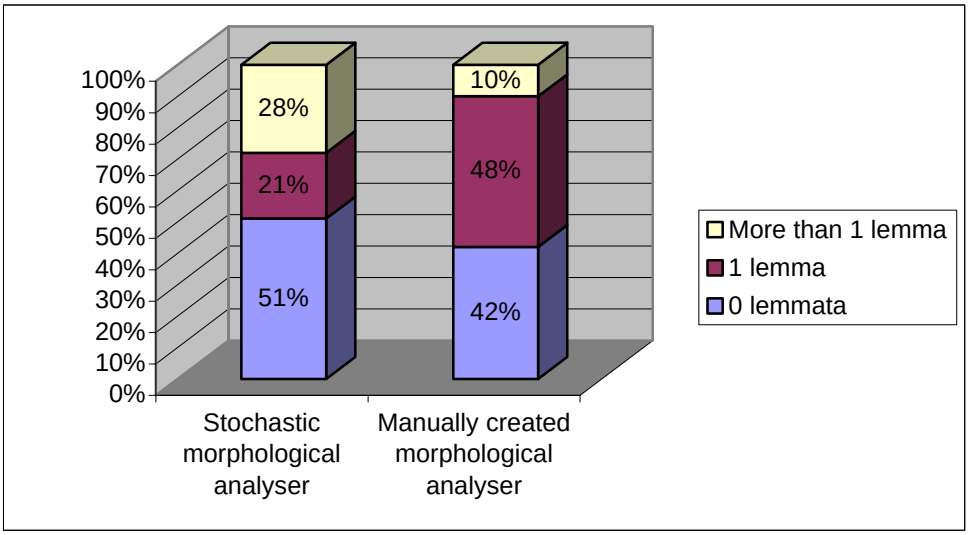


Figure 98 Comparison of the efficiency of stochastic and manually created morphological analysers

Given that the manually created morphological analyser runs so slowly, it is very important that it be used to analyse the list of word types, rather than it be applied directly to the word tokens in the corpus. The manually created morphological analyser could possibly be made to identify further possible lemmata and, thereby, leave fewer word types for which no lemmata are suggested. This would require the addition of more rules and/or more complicated rules. One could, for example, include rules which would account for inflection by vowel affection or infixation or would take into account morphophonemic alternation. However, the addition of such extra rules would cause the lemmatiser to become many times slower. The 2½ weeks that were taken to analyse 9610 word types could easily become 2½ months. The choice then is between creating a relatively simple morphological analyser that completes its analysis within certain time constraints and a more thorough morphological analyser that may take an extremely long time to complete

its analysis.

No lemmatisation system that is based solely on affix stripping can handle suppletion. Suppletive word types can only be added to the lemmatisation database manually. A morphological analyser can thus only ever be a partial solution to creating a lemmatisation database.

5.8 Homograph Separation

Certain lemma signs may be interpreted as homonymous and each homonym treated in a different article (Hausmann and Wiegand 1991: 337). Any decision concerning the number of base forms may depend on whether the dictionary is intended for encoding or decoding (Schnorr 1991: 2813 *ff.*). Whilst computer programs may facilitate with homograph separation, Kipfer (1984: 167) is of the opinion that the process cannot be fully automated electronically and that extensive human intervention is required, in particular with regard to assigning lemmata to instances of unique spelling and citations in a concordance.

Robins (1987: 56 *ff.*) identifies four criteria which may be used to identify separate words as well as separate meanings of a single word: formal semantic distinctiveness, etymology, grammatical differences, and collocational sets.

Zgusta (1971: 74) maintains that homonymy commences at the point where the speakers of a language are not able to perceive different senses as being related.

> Of course it is a pity that we have to rely on the subjective interpretations of the speakers, but we have hardly anything else on hand. And after all, a language exists to be spoken and understood, and it exists by being spoken and understood, so that intersubjective understanding of the speakers can be considered a criterion.
>
> (Zgusta 1971: 75)

Zgusta (1971: 78) recommends that only pairs with vastly different unconnected meanings be acknowledged as homonyms.

Etymology may sometimes be employed as a criterion to corroborate the intersubjective attitude of native speakers towards the semantic relatedness of a pair of words. Another language is frequently the source for one member of the pair. Occasionally both of the items are borrowed from a language in which the original forms differed. It is not possible to restrict the notion of homography to the etymological distinctiveness of a pair of items. Indeed sometimes the etymology is unknown (Zgusta 1971: 76-86).

Morton Nance (NCED) gives *colon* two entries. To the first of these he gives the English translation equivalent ‘heart’, to the second he gives ‘gut’, ‘entrail’, ‘bowel’, ‘belly’. Since both refer to internal organs one might be excused for thinking that they should be treated as a single lexeme. However **colon** (‘heart’) appears cognate with Welsh CALON (‘heart’) and Breton KALON (‘heart’, ‘centre’, ‘courage’), whereas **colon** (‘gut’, ‘entrail’, ‘bowel’, ‘belly’) appears to be a shortened form of **colodhyon** (‘bowel’) and cognate with Welsh COLUDDION (‘bowel’).

The criterion of formal grammatical difference distinguishes homographs by their different paradigms. Zgusta (1971: 81) advises that the lexicographer

should not consider a single form but should take the whole paradigm into consideration and heed all its oblique forms. If two words with the same form are distinguished by their paradigms they should be treated as homographs. In Cornish two homographs that share the base form *er* may be distinguished by their plurals.

er: 'heir'; *eryon*: 'heirs'.

er: 'temple'; *eryow*: 'temples'.

According to Zgusta (1971: 81), when there are differences between the paradigms but the base forms are identical this is referred to as partial homonymy. Total homonymy, in contrast, involves all forms of the two lexemes being identical.

Paradigmatic difference may signal a difference in part-of-speech. In Cornish, two homographs of *bos* can be distinguished as follows:

bos: 'abode', 'dwelling'; *bosow*: 'abodes', 'dwellings',

bos: be; *ov*: 'I am'; *os*: 'thou art'; *yw*: 'he/she/it is'; *on*: 'we are'; *etc*.

In the case of languages, such as Cornish, overt gender marking may be used to distinguish homographs: for example **goeth** (masculine) 'pride', and **goeth** (feminine) 'stream'.

The recognition of 'grammatical neutrals' removes the need for multiple listing of grammatical homographs. Morton Nance (NCED) gives separate entries for **gothvos** (verb) 'to know', and **gothvos** (noun) 'knowledge'. George (GKK) deals with these in a single entry (see also Zgusta 1971: 85). However this introduces an additional complexity as a number of classes can be

combined: noun and adjective, adjective and verb, adjective adverb.

Differences in the derivational series are sometimes considered to signal homography. Zgusta (1971: 86) is of the opinion a cleavage in the derivational series is strongly indicative of the bases of derivation becoming polarised in their meaning and mutual status; but Zgusta does not feel that this is a decisive criterion for distinguishing homographs.

Part-of-speech may be identified using, semantic, phonological, morphological or syntactic criteria. These criteria, however, do not necessarily achieve the same result. It is important, then, to consider which criteria are to be used and what effect this will have for the dictionary user.

According to Zgusta (1971: 250),

> The other indications of the lemma inform the user about the (usually morphological but - above all in the case of uninflected words - also the syntactic or combinatorial) class of which the entry word (i.e. the respective lexical unit) is a member. This can be indicated either by the cardinal forms of each respective paradigm, or by the number of the paradigm (according to their numeration either as generally accepted, or as numbered in the grammatical sketch appended to the dictionary), or by an abbreviation or sign (e.g. n. - noun in an Eng. dictionary), or by any other similar means. ... it is necessary to state fully and explicitly, in the forward to the dictionary, what classes and categories are indicated, and by what means. Second, the bigger the dictionary, the more imperative the necessity to indicate all eventual aberrations of the respective lexical unit from the usual paradigm, i.e. to indicate all its “irregular forms”. ... The non-existence of a form etc. (e.g. “no plural”) should also be indicated.

In the case of the bilingual dictionary, the lexicographer must establish which part-of-speech categories are present in both the languages. It is then necessary to decide which pairs of categories are to be considered equivalent (Zgusta

1971: 313).

Semantic properties are sometimes used to determine part-of-speech. Traditionally it is said that verbs denote actions (MOAZ, ‘go’; PONYE, ‘run’; DEBRY, ‘eat’; etc.); nouns denote entities (GWETHEN, ‘tree’; CATH, ‘cat’; BRE, ‘hill’; etc.); adjectives denote states (CLAF, ‘ill’; LOWEN, ‘happy’; GOCY, ‘foolish’; etc.); prepositions denote location (DAN, ‘under’; WAR, ‘on’; ADRO, ‘around’; etc.). It is sometimes said (Radford 1988: 57) that these traditional semantic criteria do not provide a reliable means of classification. For example, CAREESK (‘Exeter’) denotes a location but is a noun not a preposition; CLEVAS (‘illness’) denotes a state but is a noun not an adjective. However a more detailed set of semantic criteria emerge if this taxonomy is extended.

Thus the entities that nouns denote may be concrete or abstract. Concrete entities consist of material, physical substance and may be either animate or inanimate. Abstract entities are intangible. Furthermore, nouns may be proper or common. Proper nouns name entities whose reference is unique. In other words, a proper noun has a single specific or generalised denotation. Personal names, place names, days and months are examples of proper nouns. Common nouns denote entities that lack unique reference. Common nouns may be either countable or mass. Using this taxonomy, CAREESK (‘Exeter’) is a proper noun; and CLEVAS (‘illness’) may be classified as an abstract, uncountable common noun. Figure 99 shows the nominal system based on semantic criteria.

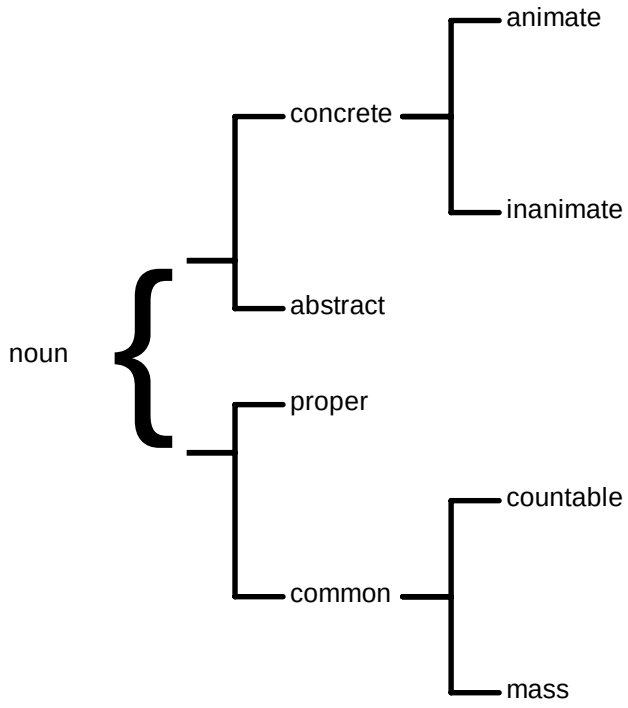


Figure 99 The nominal system based on semantic criteria

Verbs may denote one of three types of process: action, event or state. Actions may be identified by questions of the type, “What is X doing?” or “What did X do?”. Events may be identified by questions of the type, “What is happening?” or “What happened?”. States may be identified by questions of the type, “What is/was the state of the subject?”. Using this taxonomy, KERRAS (‘walk’) can be identified as an action-verb from the following attestation.

> “an Arluth Deew a **kerras** en Looar” (*Gwavas Manuscripts*: 100r)
>
> ‘the Lord God was **walk**ing in the garden’

Similarly, TRAYLYAH (‘return’) can be identified as an event-verb from the

following attestation.

> “ha tha douste che ra **traylyah**” (*Gwavas Manuscripts*: 101r)
>
> ‘and unto dust shalt thou **return**’

Finally, BOSE (‘be’) can be identified as a state-verb from the following attestation.

> “Drefan **bose** mar deake tha face” (*Gwreans an Bys*: line 564)
>
> ‘Because your face **is** so pretty’

Alternatively verb processes may be classified as material, mental, verbal, behavioural, existential or relational (Halliday 1994: 106 ff.). The material process is a process of doing; it involves an actor. Material processes may be identified by questions of the type, “What did X do?” or What did X do to Y?”. The mental process is a process of sensing; it involves a sensor and a phenomenon. The mental process may be further subdivided into processes of perception, affection and cognition. Mental processes may be identified by questions of the type, “What did X think/feel/know about Y?”. The verbal process is a process of saying. The behavioural process has no clearly defined characteristics of its own, being partly like the material process and partly like the mental process. The behavioural process represents physiological and psychological behaviour. The existential process represents that something exists or happens. The relational process is a process of being in which things are stated to exist in relation to other things. The relational process may further subdivided into processes of identification and attribution. The identification process defines something. The attributive process ascribes a quality, classification or epithet. Figure 100 shows the system of verbal

processes.

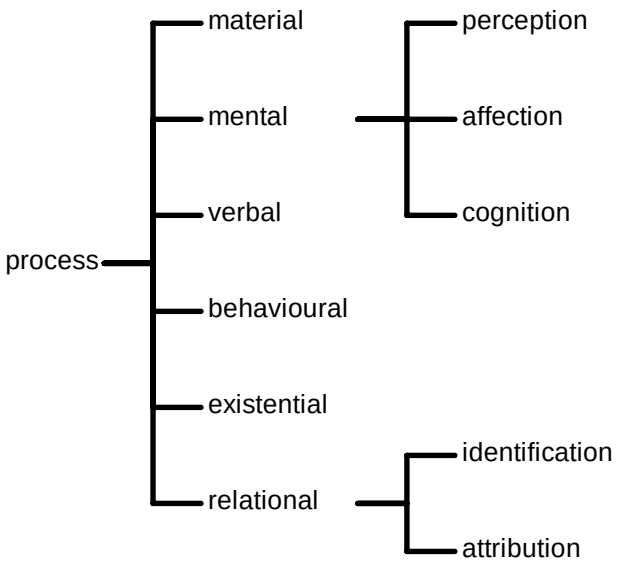


Figure 100 The system of verbal processes

From a semantic perspective, adverbs fall into several categories: circumstantial adverbs, adverbs of degree, conjunctive adverbs, modal/disjunctive adverbs, and interpersonal adverbs. Figure 101 shows the adverbial system based on semantic criteria.

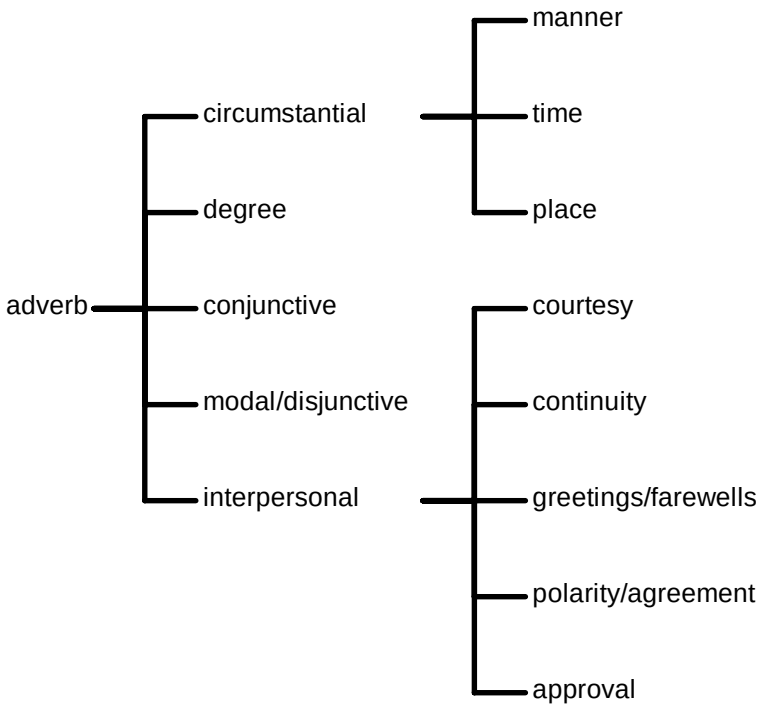


Figure 101 The adverbial system from a semantic perspective

Circumstantial adverbs are single words marking the circumstances of the verbal process and fall into the categories, manner, time and place. Adverbs of manner answer the question "How?". Thus DYSON ('silently') can be identified as an adverb of manner from the following attestation.

> "tan henna theworthef vy **dyson"** (*Origo Mundi*: line 207)
>
> 'take that from me **silently'**

Adverbs of time answer the question "When?" Thus LEBMYN ('now') can be identified as an adverb of time from the following attestation.

> des thymma ve **lebmyn** (*Gwreans an Bys*: line 2241)
>
> come to me **now**

Adverbs of place answer the question “Where?”. Thus ENA (‘there’) can be identified as an adverb of place from the following attestation.

> “ty a the eynda hag **ena** pregoth a wra” (*Resurrexio Domini*: line 2458)
>
> ‘thou shalt go to India and **there** shalt preach’

Adverbs of degree indicate the degree, extent or intensity. Thus FEST (‘very’) can be identified as an adverb of degree from the following attestation.

> me an knouk **fest** dybyte (*Passio Domini*: line 2091)
>
> I will beat him **very** cruelly

Conjunctive adverbs provide a connective link between the present clause and the preceding one. Thus BETEGYNS (‘nevertheless’) can be identified as a conjunctive adverb from the following attestation.

> “avel fol y an scornye hag an gweska fest yn tyn **betegyns** ger ny gewsy” (*Pascon Agan Arluth*: stanza 114)
>
> ‘They mocked him like a madman and beat him most cruelly. **Nevertheless** he did not speak a word.’

Modal adverbs express various aspects of the speaker’s perspective on the sentence. Thus CERTUS (‘certainly’) can be identified as a modal adverb from the following attestation.

> “**certus** rag the gerense syr urry a fyth lethys” (*Origo Mundi*: line 2122)
>
> ‘**Certainly,** for your love, Sir Uriah shall be put to death.’

Interpersonal adverbs fall into several categories: courtesy, continuity markers, greeting/farewells, polarity/agreement, and approval.

Courtesy adverbs express politeness towards the addressee. Thus

GRANTMERCI (‘thank you’) can be identified as a courtesy adverb from the following attestation.

> “Grantmerci syr iustis” (*Resurrexio Domini*: line 95)
>
> ‘Thank you, Sir Justice.’

Continuity markers signal that a response to a previous utterance is about to be provided. Thus LEBBEN (‘now’) can be identified as a continuity marker from the following attestation.

> “**Lebben** an hagar-breeve o mouy foulze a vell onen vethell an Bestaz an gweale” (Gwavas Manuscripts: 99v)
>
> ‘**Now** the serpent was more subtle than any beast of the field’

Greeting/farewell adverbs serve as salutations. Thus HOW (‘hello’) can be identified as a greeting adverb from the following attestation.

> “**how** ty geyler dus yn rak” (*Resurrexio Domini*: line 1989)
>
> ‘**Hello**, thou jailer, come forth!’

Polarity and agreement responses serve to confirm or deny a previous utterance. Thus YEA (‘yes’) can be identified as an agreement adverb from the following attestation.

> “**Yea** gwra thym indella” (*Gwreans an Bys*: line 845)
>
> ‘**Yes**, do so for me.”

Approval formulae serve to express approval or disapproval. Thus DAR (‘alas’) can be identified as an approval adverb from the following attestation.

> “**Dar** marow yu syr urry” (*Origo Mundi*: line 2217)
>
> ‘**Alas**, Sir Uriah is dead.’

From a semantic perspective, pronouns fall into three types: substantive, determinative and numerative. Substantive pronouns answer the question ‘Who?’ or ‘What?’. Determinative pronouns answer the question ‘Whose X?’, ‘Which X?’ or ‘What kind of X?’. Numerative pronouns answer the question ‘How many/much X?’. Thus in the attestation, “**me** re goskes” (*Resurrexio Domini*: line 511) - ‘**I** have slept’, ME (‘I’) can be identified as a substantive pronoun. In the attestation, “**ow** feryl” (*Origo Mundi*: line 197) – ‘**my** peril’, OW (‘my’) can be identified as a determinative pronoun. And in the attestation, “rak kuthe **myns** us formyys” (*Origo Mundi*: line 22) – ‘to cover **all** that is created’, MYNS (‘all’) can be identified as a numerative pronoun.

The semantic relations into which items enter provide further evidence for their part-of-speech. Verbs enter into the semantic relations of troponymy and entailment with one another. A troponym is a verb expressing a specific manner elaboration of another verb. For example, GWAYA (‘move’) has the troponyms KERRAS (‘walk’), POONIA (‘run’) and LEBMAL (jump, leap). In the case of entailment, a verb X entails Y if X cannot be done unless Y is, or has been done. For example, DEVINA (‘waking’) entails CUSKA (‘sleeping’).

Nouns enter into the semantic relations of hyponymy and antonymy with one another. Hyponymy involves membership of a class. In other words, X is a hyponym of Y if X is a (kind of) Y . For example, DAR (‘oak’), ON (‘ash’) and FAWE (‘beech’) are all hyponyms of GWETHAN (‘tree’). Concrete nouns enter into the semantic relation of meronymy with one another.

Meronymy is concerned with the relationship of part to whole. In other words, X is a meronym of Y if X is a part of Y . For example, SCORAN (‘branch’), BARRAN (‘twig’) and DELKIAN (‘leaf’) are all meronyms of GWETHAN (‘tree’). Some nouns enter into a member-collection relationship. In other words, X is a member of Y, and Y is a collection of Xs. For example, PYSK (‘a fish’) is a member of HEAZ (‘a shoal’). Some nouns enter into a portion-mass relationship. In other words, X, a countable noun, is a unit of measurement or division of a mass noun, Y. For example, BADNA (‘a drop’) is a portion of LIDN (‘liquid’).

The semantic relation of antonymy is particularly common between adjectives. Some adjectives can enter into the semantic relation of pertainymy. Adjectives that are pertainyms do not have antonyms and are usually defined by such phrases as “of or pertaining to”. A pertainym can point to another pertainym or a noun. For example, HAGAR (‘ugly’) is a pertainym of HACTER (‘ugliness’).

In the case of discontinuous lexemes, from a syntactic point of view, the elements that make up the lexeme are each allocated their part-of-speech. From a semantic perspective, however, a discontinuous lexeme may be allocated a single part-of-speech. The phrasal verbs found in Germanic languages are a case in point. GWYN VYS is an example of a Cornish discontinuous lexeme in which the elements GWYN and BYS (‘happy’) may be interrupted by a possessive pronoun, as can be seen from the following attestations.

“guyn vys” (*Passio Domini*: line 156) ‘happy’

“guyn ou bys” (*Passio Domini*: line 3193) ‘happy I’

“guyn the vys” (*Passio Domini*: line 156) ‘happy thou’

“guyn y vys” (*Passio Domini*: line 122) ‘happy he’

“gwyn agan bys” (*Pascon Agan Arluth*: stanza 4) ‘happy we’

“guyn aga beys” (*Resurrexio Domini*: line 279) ‘happy they’

Another example is DEAN AN PUSKES (‘fisherman’) which is found attested as “dean bodgack an puscas” (*William Bodinar’s Letter*) ‘a poor fisherman’.

Structural criteria for determining word class categories include phonological, morphological and syntactic criteria. Radford (1988) is of the opinion that morphological and syntactic criteria are a far more reliable than semantic criteria for determining word-level categories.

In the case of Cornish, phonological criteria for determining word class categories are based on initial mutations of consonants. Certain initial mutations help to identify certain nouns, verbs and adjectives.

Feminine singular nouns are lenited after the definite article, AN and the indefinite article, UN. Masculine plural nouns are lenited after the definite article, AN. Figure 102 shows some examples of nominal mutation.

Radical			**Initial Lenition**		
"benen" (f)	(*Passio Domini*: line 768)	'woman'	"an venen"	(*Passio Domini*: line 516)	'the woman'
"gwreag" (f)	(*Gwreans an Bys*: line 876)	'wife'	"un wreag"	(*Gwreans an Bys*: line 1449)	'a/one wife'
"tus" (m. pl.)	(*Resurrexio Domini*: line 833)	'men'	"an dus"	(*Resurrexio Domini*: line 972)	'the men'

Figure 102 Examples of nominal mutation

Verbs are lenited after the affirmative particle, A, the negative particles, NA(G) and NY(NS), the optative particle, RE, and the perfective particle, RE. Figure 103 shows some examples of verbal lenition.

Radical			***Initial Lenition***		
"crys thym"	(*Resurrexio Domini*: line 965)	'believe me'	"me a grys"	(*Passio Domini*: line 3078)	'I believe'
"cowse"	(*Gwreans an Bys*: line 164)	'say'	"na gowse"	(*Gwreans an Bys*: line 171)	'Don't say'
"guraf"	(*Origo Mundi*: line 25)	'I do'	"my ny wraf"	(*Passio Domini*: line 901)	'I do not'
"tryge"	(*Passio Domini*: line 808)	'abide'	"re drygas"	(*Passio Domini*: line 805)	'have abided'

Figure 103 Examples of verbal lenition

Verbs are provected after the present participle particle, OW(TH*)*, and the conditional particles, MAR(A)(S) and A. Figure 104 shows some examples of verbal provection.

Radical			Initial Provection		
“guerthe”	(*Passio Domini*: line 1108)	‘sell’	“ow querthe”	(*Passio Domini*: line 1520)	‘selling’
“gallaf”	(*Gwreans an Bys*: line 1709)	‘I can’	“mara callaf”	(*Gwreans an Bys*: line 1442)	‘if I can’
“gallus”	(*Gwreans an Bys*: line 357)	‘power /ability’	“a callen”	(*Gwreans an Bys*: line 785)	‘if we could’

Figure 104 Examples of verbal provection

Verbs undergo the mixed mutation after the particle, Y(TH*)*. Figure 105 shows some examples of verbal mixed mutation.

Radical			Mixed Mutation		
“mennaf”	(*Passio Domini*: line 232)	‘I want’	“y fynna”	(*Origo Mundi*: line 17)	‘I will’
“gwrens tus”	(*Gwreans an Bys*: line 2168)	‘Let men’	“y wrens”	(*Pascon Agan Arluth*: stanza 39)	‘they would’

Figure 105 Examples of verbal mixed mutation

Adjectives are lenited after feminine singular nouns and after masculine plural nouns. Adjectives are also lenited after ONEN when used as a pronoun and referring to feminine singular nouns. Figure 106 shows some examples of adjectival lenition.

Radical			Initial Lenition		
“mas”	(*Resurrexio Domini*: line 2487)	good	“benen (f.) vas”	(*Resurrexio Domini*: line 1697)	‘good wife’
“bras”	(*Passio Domini*: line 171)	great	“tus (m.pl.) vras”	(*Passio Domini*: line 790)	‘great men’
“tek”	(*Passio Domini*: line 36)	pretty	“onan dek”	(*Passio Domini*: line 2840)	‘a pretty one’

Figure 106 Examples of adjectival lenition

Brown (1984: 17, 59; 1993: 12, 56) maintains that adjectives are lenited after dual nouns of both genders. However this is not attested in the corpus. Instead

one find “thulef claf”, ‘a pair of leprous hands’ (*Passio Domini*: line 2697) in which “claf” remains unlenited.

Adjectives undergo the mixed mutation when following the adverbial particle, YN. Figure 107 shows some examples of adjectival mixed mutation.

Radical			**Mixed Mutation**		
“da”	(*Passio Domini*: line 121)	‘good’	“yn ta”	(*Passio Domini*: line 156)	‘well’
“gulan”	(*Passio Domini*: line 836)	‘clean’	“yn wlan”	(*Passio Domini*: line 2405)	‘cleanly’
“beu”	(*Passio Domini*: line 115)	‘alive’	“yn few”	(*Resurrexio Domini*: line 1442)	‘alive’

Figure 107 Examples of adjectival mixed mutation

Smith (1984: 38) observes that initial mutations are frequently found to be missing in the corpus of Cornish and he gives the following frequencies of missed mutations.

Pascon Agan Arluth	1 in every 74 lines.
Origo Mundi	1 in every 21½ lines.
Passio Christi	1 in every $11^{1}/_{3}$ lines.
Resurrexio Domini	1 in every $10^{2}/_{3}$ lines.
Beunans Meriasek	1 in every $9^{2}/_{3}$ lines.
Gwreans an Bys	1 in every 48 lines.

Smith notes that *Pascon Agan Arluth* follows the rules of mutation most accurately followed by *Gwreans an Bys*. Smith maintains that in the *Ordinalia* (*Origo Mundi, Passio Christi* and *Resurrexio Domini*) the phrases *a pup, the pup* and *war pup* are never found to be mutated, whereas in *Gwreans an Bys,* they are always mutated. Smith concludes that omission of mutations is,

therefore, merely scribal. Smith, however, is not quite accurate in his assertion that *a pup* and *war pup* are never found to be mutated in the *Ordinalia*. In the *Ordinalia*, WAR + PUP is attested only once with initial mutation of <P> to <B>, “war bup” (*Origo Mundi*: line 77). In the case of A + PUP, there are 15 attestations of “a pup” (*Passio Domini*: lines 416, 477, 844, 865, 2307, 2418, 2937, 3056; *Resurrexio Domini*: line 1600, 1652, 1671, 1742, 1757, 2346, 2558,) and 1 attestation of the mutated form, *a bop* (*Passio Domini*: line 838). In *Gwreans an Bys*, THE + PUP is not attested. Nevertheless Smith’s suggestion that the observance or omission of initial mutations is scribal is plausible. It must be concluded that idiosyncrasies in scribal performance mean that it is an unreliable indicator of part-of-speech.

Two types of morphological criterion may be employed to determine part-of-speech: inflection and derivation. Radford (1988) points out that certain types of inflection attach only to specific categories. Thus individual categories are distinguished by the range of inflections that they permit. Initially words may be divided into two categories: those that permit inflection and those that do not. Figure 108 shows the Cornish inflection system.

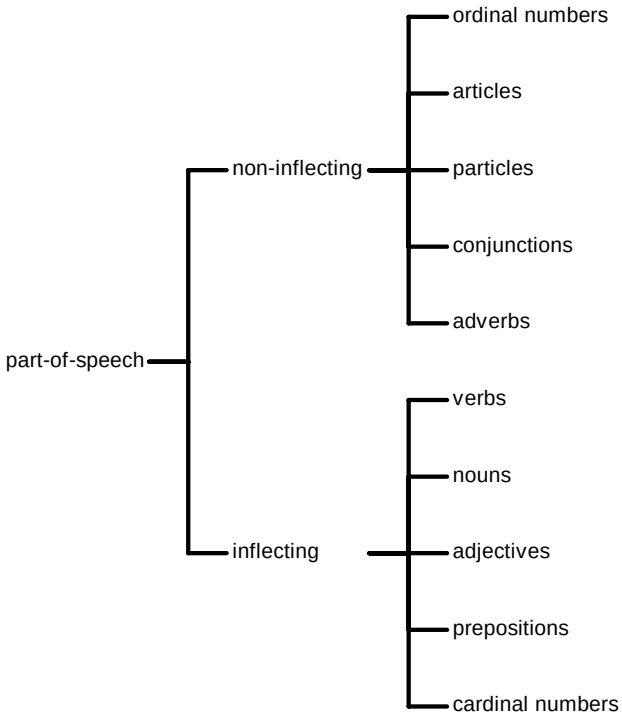


Figure 108 The Cornish inflection system

Non-inflecting categories include ordinal numbers, articles, particles, conjunctions and adverbs. Those categories which permit inflection include verbs, nouns, adjectives, prepositions and certain cardinal numbers.

In the case of Cornish verbs, inflection is marked by a combination of suffixes and vowel affection of the stem. The uninflected stem is used for the third person singular of the present tense and for the second person singular of the imperative. Inflections mark the non-finite forms of the verb. The participle is formed by adding -ES to the root. The infinitive of the verb can take several forms; it may consist of only the stem or may take one of the following suffixes: -A, -YA, -E, AL, -EL, -AS, -ES, -Y. Finite forms are inflected for person, number, tense and mood. Figure 109 shows the inflections of

the Cornish verb, CARA.

		MOOD					
		Indicative				Subjunctive	Imperative
		TENSE					
		Present	Imperfect	Preterite	Pluperfect		
Person & Number	1s.	*caraf*	*caren*	*kerys*	*carsen*	*kyryf*	
	2s.	*keryth*	*cares*	*kersys*	*carses*	*kyry*	*car*
	3s.	*car*	*care (cara)*	*caras*	*carse*	*caro*	*cares/carens*
	1p.	*keryn*	*caren*	*kersyn*	*carsen*	*kyryn*	*caren*
	2p.	*carough*	*careugh*	*carsough*	*carseugh*	*kyreugh*	*careugh*
	3p.	*carons*	*carens*	*carsons(ans)*	*carsens*	*carons*	*carens*
	0	*carer(yr)*		*caras*			

Figure 109 Inflections of the verb CARA

Prepositions in Cornish are inflected for number and person. Figure 110 shows the inflections of the preposition, YN.

1s.	*ynnof*	'in me'
2s.	*ynnos*	'in you'
3s.m.	*ynno*	'in him'
3s.f.	*ynny*	'in her'
1p.	*ynnon*	'in us'
2p.	*ynnough*	'in you'
3p.	*ynna*	'in them'

Figure 110 Inflections of the preposition YN

Comparative and superlative forms of adjectives are marked by the suffix, -A. Figure 111 shows the inflections of the adjective, UHEL.

"uhel" (*Passio Domini*: line 1716) 'high'	"uhella" (*Passio Domini*: line 2189) 'higher'	"an ughella" (*Gwreans an Bys*: line 39) 'the highest'

Figure 111 Inflections of the adjective, UHEL

Most of the time adjectives are not found to be inflected for number. One adjective only is commonly inflected for number; ARAL has the plural form, *erel*. However Lhuyd (AB: 243-4) gives *dyon* as the plural form of DIU

(‘black’). Lhuyd’s (AB: 243) assertion that a masculine adjective containing the vowel <Y> may be made feminine by changing the <Y> to <E> does not appear to be supported by attestation in the corpus.

Cornish nouns are inflected for number and can take up to five forms: singular, plural, collective, singulative and dual. Figure 112 shows some examples of Cornish nominal inflection.

Singular		**Plural**		**Collective**		**Singulative**		**Dual**	
		“dowrow” *Gwreans an Bys*	‘waters’, ‘water-places’	“dour” (*Origo Mundi*: line 1833)	‘water’	“dowren”	‘a water-place’		
		“sterennow”	‘stars’	“steyr” (*Pascon Agan Arluth*: stanza 211)	‘stars’	“sterran” (AB: 121)	‘a star’		
“dar” (VC)	‘an oak’	“derow” (*Origo Mundi*: line 1010)	‘oaks’			“derowen”	‘an oak’		
“daves” (*Origo Mundi*: line 127)	‘a sheep’	“devidgyow” (*Gwreans an Bys*: line 1068)	‘sheep’	“deves” (*Passio Domini*: line 894)	‘sheep’				
“ger” (*Passio Domini*: line 1431)	‘a word’	“geryow” (*Passio Domini*: line 2468), “gerennow” (*Beunans Meriasek*: line 2964)	‘words’	“ger” (*Passio Domini*: line 1431)	‘an utterance’	“geren”	‘a single word’		
“gueder” (AB: 18, 175)	‘a glass vessel’	“gwedrennow”, “gwedrow”	‘drinking glasses’	“gueder” (AB: 18, 175)	‘glass’	“guedran” (AB: 242)	‘a drinking glass’		
“huneys”	‘a sleep’			“hun” (*Origo Mundi*: line 1921)	‘sleep’				
“luef” (*Passio Domini*: line 2747)	‘a hand’	“lufyow”	‘hands’					“dyulef” (*Passio Domini*: line 2375)	‘a pair of hands’
“men” (*Resurrexio Domini*: line 400)	‘a stone’	“menow”	‘stones’	“myn” (*Resurrexio Domini*: line 401)	‘stones’				
“tros” *Origo Mundi*: line 262	‘a foot’			“treys” (*Passio Domini*: line 474)	‘feet’			“deutros”	‘a pair of feet’

Figure 112 Examples of Cornish nominal inflection

Some, but not all, of the cardinal numbers are inflected for gender. Figure 113 shows some examples of Cornish cardinal numeric inflection.

masculine	feminine	
un	*un*	‘one’
deu	*dyw*	‘two’
try	*tyr*	‘three’
peswar	*pedwar*	‘four’
pymp	*pymp*	‘five’

Figure 113 Examples of Cornish cardinal numeric inflection

Derivational morphological criteria may also serve to indicate word class. By the addition of suffixes, nouns may be derived from adjectives or verbs, and adjectives may be derived from nouns.

Abstract nouns are derived from adjectives by adding -TER after a voiceless consonant and -DER after other letters. Figure 114 shows examples of nouns derived from adjectives by the addition of –TER or -DER

Adjective			Abstract Noun		
“whek”	‘sweet’	(*Passio Domini*: line 35)	“whekter”	‘sweetness’	(*Origo Mundi*: line 359)
“da”	‘good’	(*Pascon Agan Arluth*: stanza 37)	“dader”	‘goodness’	(*Passio Domini*: line 3096)
“gwan”	‘weak’	(*Pascon Agan Arluth*: stanza 205)	“gwander”	‘weakness’	(*Pascon Agan Arluth*: stanza 68)

Figure 114 Nouns derived from adjectives by the addition of –TER or -DER

Alternatively nouns may be derived from adjectives by adding -ETH or -NETH. Figure 115 shows some examples of nouns derived from adjectives by the addition of -(N)ETH. Figure 115 shows some examples of nouns derived from verbs the addition of -(N)ANS.

Adjective			**Abstract Noun**		
“fol”	‘foolish’	(*Resurrexio Domini*: line 2182)	“folneth”	‘folly’	(*Resurrexio Domini*: line 961)
“cosel”	‘calm’	(*Origo Mundi*: line 2074)	“cosoleth”	‘calm’	(*Origo Mundi*: line 518)
“goky”	‘stupid’	(*Passio Domini*: line 1662)	“gokyneth”	‘stupidity’	(*Origo Mundi*: line 1512)

Figure 115 Nouns derived from adjectives by the addition of -(N)ETH

Abstract nouns are derived from verbs by adding -ANS or -NANS.

Verb			Abstract Noun		
“crygy”	‘believe’	(*Resurrexio Domini*: line 1088)	“crygyans”	‘belief’	(*Passio Domini*: line 1813)
“bewe”	‘live’	(*Origo Mundi*: line 62)	“bewnans”	‘life’	(*Origo Mundi*: line 63)
“dysquethas”	‘declare’	(*Origo Mundi*: line 1439)	“dysquythyans”	‘declaration’	(*Origo Mundi*: line 1733)

Figure 116 Nouns derived from verbs the addition of -(N)ANS.

An agentive noun may be derived from a verb by adding -OR. Figure 117 shows some examples of agentive nouns derived from verbs by the addition of –OR.

Verb			**Agentive Noun**		
“pystry”	‘work magic’	(*Passio Domini*: line 1765)	“pystryor”	‘sorcerer’	(*Passio Domini*: line 1767)
“ty”	‘cover’	(*Origo Mundi*: line 2490)	“tyor”	‘tiler’	(*Origo Mundi*: line 2411)

Figure 117 Agentive nouns derived from verbs by the addition of -OR

Alternatively agentive nouns may be derived from verbs by adding -YAS. Figure 118 shows some examples of agentive nouns derived from verbs by the

addition of –YAS.

Verb			Agentive Noun		
"guythe"	'keep'	(*Passio Domini*: line 10)	"gwythyas"	'keeper'	(*Origo Mundi*: line 692)
"selwel"	'save'	(*Passio Domini*: line 2953)	"sylwyas"	'saviour'	(*Passio Domini*: line 252)

Figure 118 Agentive nouns derived from verbs by the addition of -YAS

Adjectives are derived from nouns by the addition of -EK. Figure 119 shows some examples of adjectives derived from nouns by the addition of –EK.

Noun			Adjective		
"gallos"	'power'	(*Origo Mundi*: line 1214)	"gallosek"	'powerful'	(*Resurrexio Domini*: line 752)
"lowene"	'joy'	(*Passio Domini*: line 574)	"lowenek"	'joyful'	(*Resurrexio Domini*: line 1333)
"whans"	'desire'	(*Origo Mundi*: line 1806)	"whansek"	'desirous'	(*Passio Domini*: line 37)

Figure 119 Adjectives derived from nouns by the addition of -EK

Syntactic evidence for part-of-speech categories is concerned with distribution, the set of possible sentence positions in which an item can occur. More specifically, any word plays the part of one of the elements modifier, head or qualifier in a nominal or adjectival group; particle, auxiliary or lexical in a verbal group; opener, prepend or completive in a prepositional group or is an adverb or conjunction. Figure 120 shows a syntactic parse of the sentence "mab deu a tremyn an beys annotho del yu scryfys yn lyfryow yn lyes le" (*Passio Domini*: line 747), 'The son of God will pass from the world as it is written of him in books in many places.'

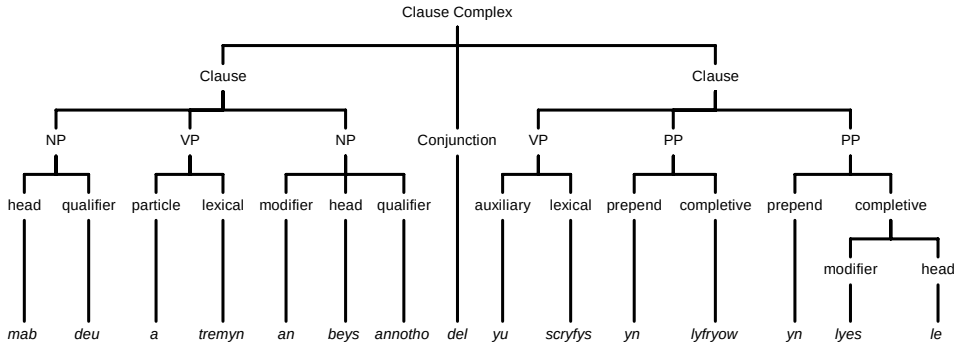


Figure 120 Possible sentence positions in which lexical items can occur

The syntactic criteria for identifying nouns are as follows. A nominal phrase always has a noun as its head. A noun may also serve as genitive qualifier in a nominal phrase. In a prepositional phrase, a noun may serve as the completive. Figure 121 shows the syntactic environments in which nouns occur.

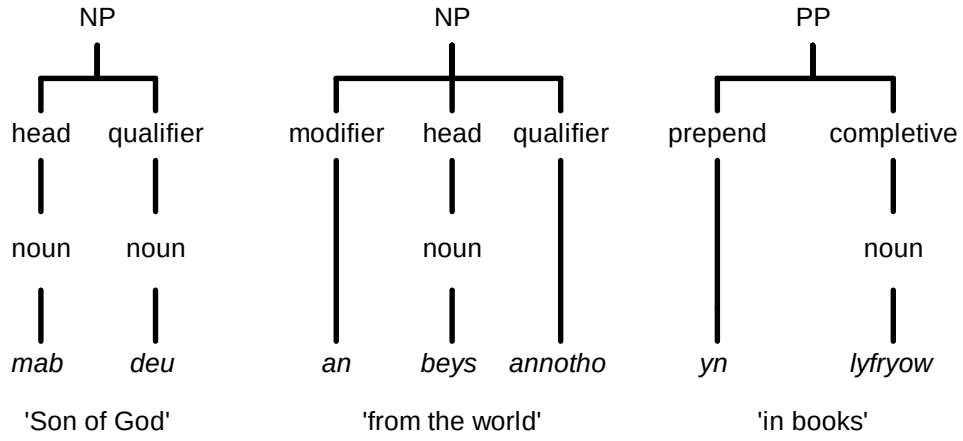


Figure 121 Syntactic environments in which nouns occur

From a syntactic perspective, there are five types of pronoun: independent,

suffixed, infixed and possessive (see Figure 122).

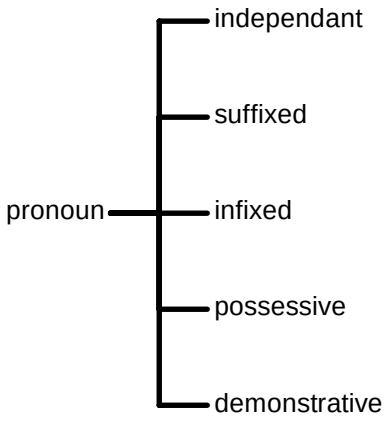


Figure 122 The Cornish pronominal system

The independent pronoun may serve as the Subject of the Clause in a periphrastic construction or as a qualifier. Thus ME/MY is an independent pronoun in the attestation, “me re goskes pos” (*Resurrexio Domini*: line 511), ‘I have slept heavily’; and in the attestation, “kepar ha my” (*Origo Mundi*: line 2350), ‘like me’. Figure 123 shows the syntactic environments in which independent pronouns occur.

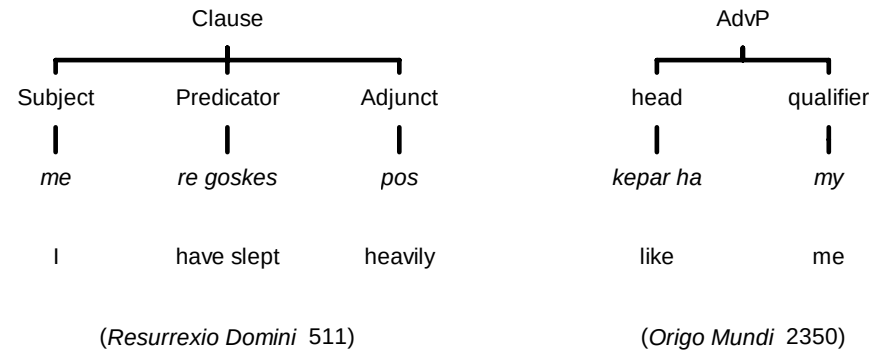


Figure 123 Syntactic environments in which independent pronouns occur

The suffixed pronoun follows the Predicator and may serve as either the Subject or the Complement of the Clause. Thus –VY is the subject in the attestation, “ny welaf vy” (*Resurrexio Domini*: line 1962), ‘I do not see’; and is the object in the attestation, “guyth vy” (*Resurrexio Domini*: line 1564), ‘preserve me’. Figure 124 shows the syntactic environments in which suffixed pronouns occur.

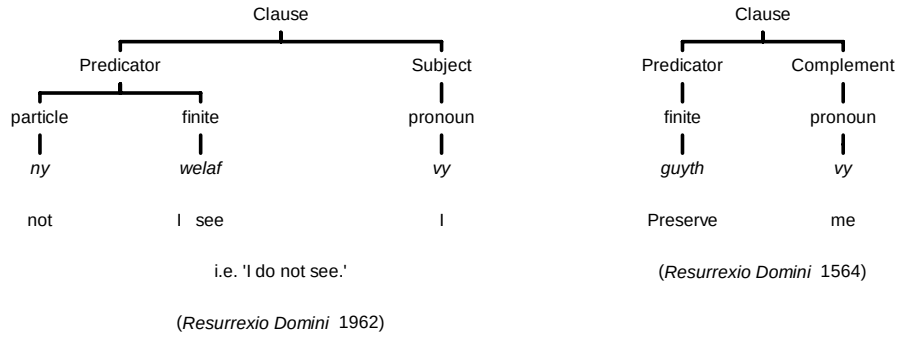
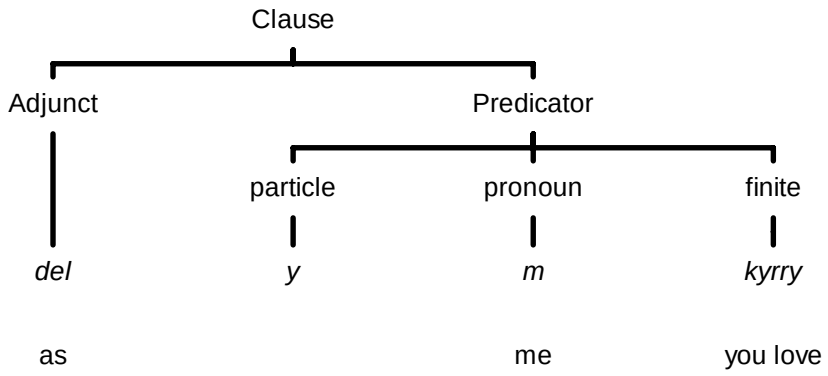


Figure 124 Syntactic environments in which suffixed pronouns occur

The infixed pronoun occurs as an element within the verbal phrase, between the particle and finite. Thus M is an infixed pronoun in the attestation, “del ym

kyrry” (*Origo Mundi*: line 2403), ‘as you love me’. Figure 125 shows the syntactic environment in which infixed pronouns occur.

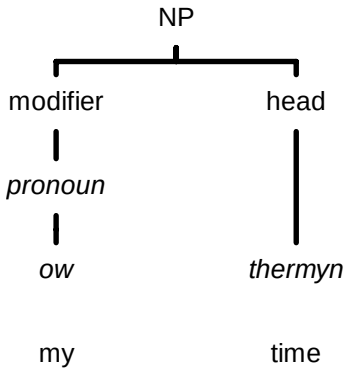


i.e. 'As you love me.'

(*Origo Mundi* 2403)

Figure 125 Syntactic environment in which infixed pronouns occur

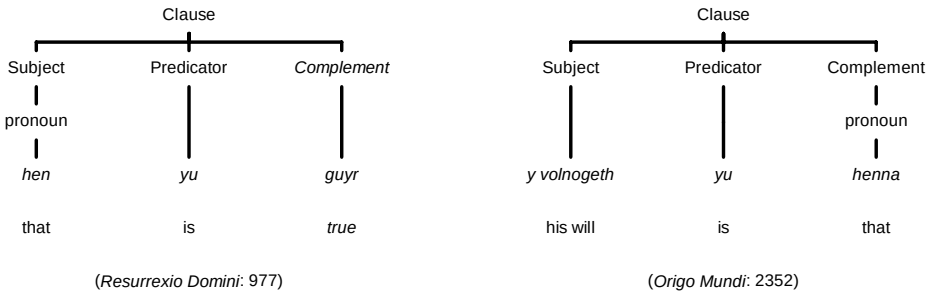
The possessive pronoun serves as a modifier in a nominal phrase. Thus OW is a possessive pronoun in the nominal phrase, “ow thermyn” (*Origo Mundi*: line 2344), ‘my time’. Figure 126 shows the syntactic environment in which possessive pronouns occur.



(*Origo Mundi* 2344)

Figure 126 Syntactic environment in which possessive pronouns occur

The demonstrative pronoun may serve as the Subject or the Complement of a Clause. Thus HEN is a demonstrative pronoun in the attestation, “hen yu guyr” (*Resurrexio Domini*: line 977) ‘that is true’; and HENNA is a demonstrative pronoun in the attestation, “y volnogeth yu henna” (*Origo Mundi*: line 2352), ‘his will is that’. Figure 127 shows the syntactic environment in which demonstrative pronouns occur.



(*Resurrexio Domini*: 977) (*Origo Mundi*: 2352)

Figure 127 Syntactic environment in which demonstrative pronouns occur

The syntactic criteria for identifying verbs are as follows. A verbal phrase

always has a verb as its head. A verb may serve as the completive in a prepositional phrase. The past participle of the verb may serve as a qualifier in a nominal phrase. Figure 128 shows the syntactic environments in which verbs occur.

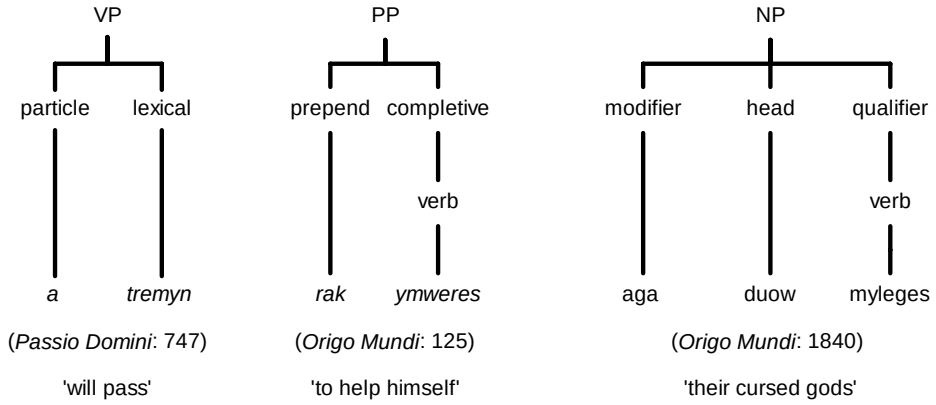


Figure 128 Syntactic environments in which verbs occur

The syntactic criteria for identifying adjectives are as follows. An adjectival phrase always has an adjective as its head. An adjective may serve as a modifier or a qualifier in a nominal phrase. An adjective may serve as a completive in a prepositional phrase. Figure 129 shows the syntactic environments in which adjectives occur.

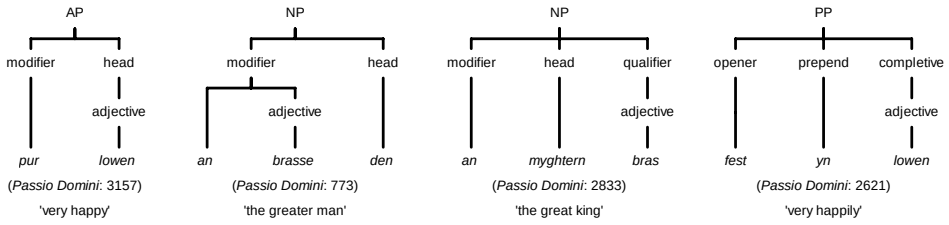


Figure 129 Syntactic environments in which adjectives occur

Adverbs may be classified as circumstantial adverbs, adverbs of degree or sentential/interclausal adverbs. A circumstantial adverb is a single word that may serve as an adjunct within the clause or as the head of an adverbial phrase. Figure 130 shows examples of circumstantial adverbs serving as Adjuncts. Figure 131 shows a circumstantial adverb as head of an adverbial phrase.

Adjunct	**Subject**	**Predicator**	**Complement**	
“ena	why	a gyf	asen”	(*Passio Domini*: line 176)
‘There	you	will find	a donkey.’	

Predicator	**Complement**	**Complement**	**Adjunct**	
“tan	henna	theworthef vy	dyson”	(*Origo Mundi*: line 207)
‘Take	that	from me	silently.’	

Figure 130 Circumstantial adverbs serving as Adjuncts

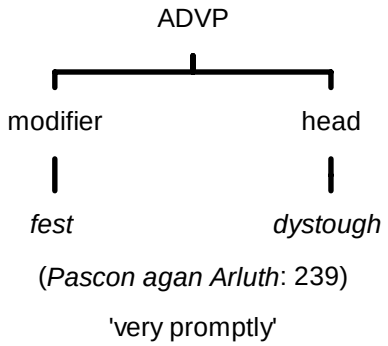
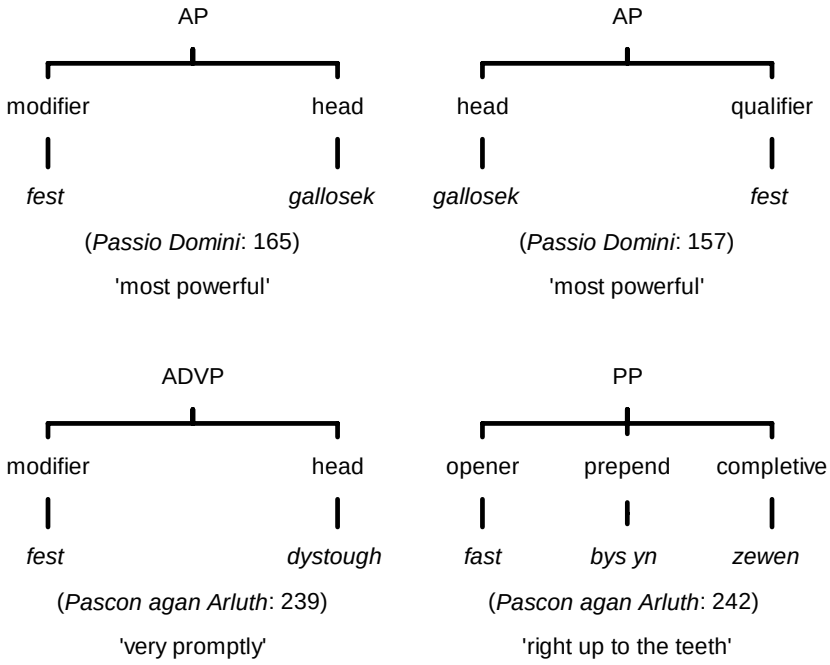


Figure 131 Circumstantial adverb as head of adverbial phrase

An adverb of degree may serve as a modifier or a qualifier in an adjectival or adverbial phrase, as an opener in a prepositional phrase, or as an adjunct within the clause. Figure 132 shows the syntactic environments in which

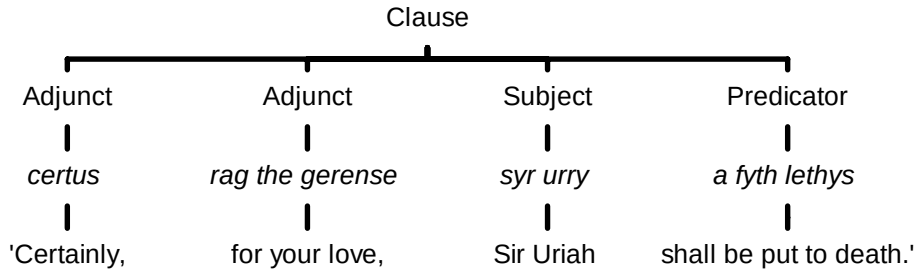
adverbs of degree may occur.



Subject[vocative]	**Complement**	**Predicator**	**Adjunct**	
"a pylat	wolcom	os	fest"	(*Resurrexio Domini*: line 1811)
'O Pilate!	welcome	thou art	very'	

Figure 132 Syntactic environments in which adverbs of degree may occur

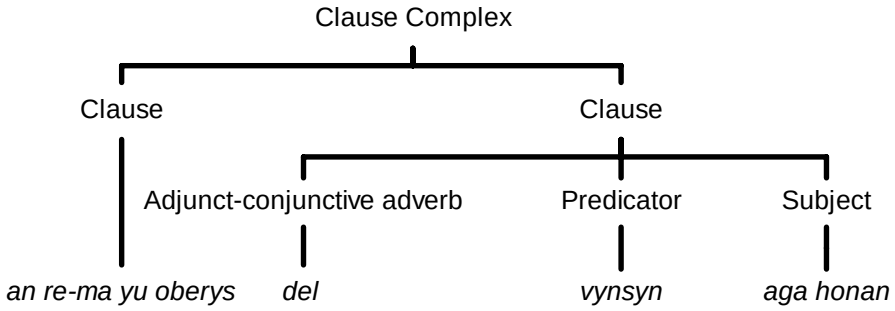
A sentential adverb is a single word that serves an adjunct within the clause. Thus CERTUS is a sentential adverb in the attestation, **"certus** rag the gerense syr urry a fyth lethys" (*Origo Mundi*: line 2122), **'Certainly,** for they love, Sir Uriah shall be put to death.' And DAR is a sentential adverb in the attestation, "**Dar** marow yu syr urry" (*Origo Mundi*: line 2217), '**Alas**! Sir Uriah is dead.' Figure 133 shows the sentential adverb, CERTUS serving as an adjunct within the clause.



(*Origo Mundi*: 2122)

Figure 133 Sentential adverb serving as an adjunct within the clause

A conjunctive adverb is distinguished from other sentential adverbs in that it provides a connective link between a pair of clauses. Figure 134 shows the conjunctive adverb, DEL, linking a pair of clauses.



(*Origo Mundi*: 11-12)

Figure 134 Conjunctive adverb linking two clauses

The syntactic criteria for identifying prepositions are as follows. A prepositional phrase always has a preposition as its head. A preposition may

also serve as a qualifier in a nominal phrase. Figure 135 shows the syntactic environments in which prepositions occur.

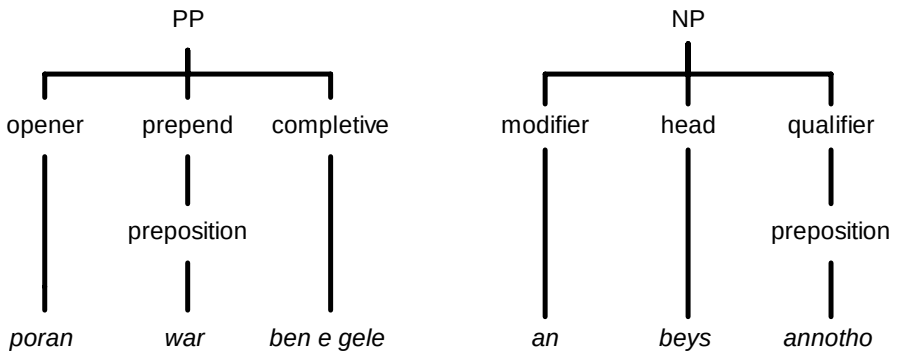


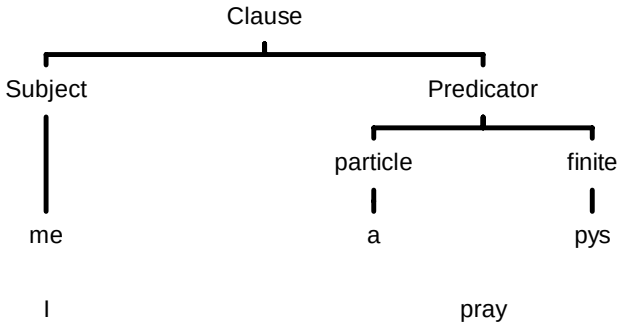
Figure 135 Syntactic environments in which prepositions occur

The syntactic criteria for identifying verbal particles and auxiliaries are as follows. The verbal phrase may be either pro-drop or periphrastic. When the verbal phrase is pro-drop, the only obligatory item in the verbal phrase is the finite element. The present participle particle follows the finite element; other particles precede the finite element. Figure 136 shows verbal particles and auxiliaries in pro-drop environments.

particle	finite	present participle particle	auxiliary	lexical	
	“guraf” ‘I do’				(*Origo Mundi*: line 1275)
	“gruaff ‘I do			routia” break’	(*Beunans Meriasek*: line 2368)
“y	lavaraf” ‘I say’				(*Origo Mundi*: line 7)
“y	fons ‘they were	ow		kronkya” beating’	(*Pascon Agan Arluth*: stanza 132)
“yth	esaf ‘I am	ow	pose being	gorthys” put’	(*Gwreans an Bys*: line 2125)
“nyns ‘not	esos you are	ou		attendya” considering’	(*Beunans Meriasek*: line 848)

Figure 136 Verbal particles and auxiliaries in pro-drop environments

The periphrastic from of the verb phrase requires a subject (see Figure 137).



(*Passio Domini*: 27)

Figure 137 The periphrastic verb phrase

In the periphrastic form of the verb phrase, the finite item is always preceded by its particle (see Figure 138).

particle	**finite**	**lexical**	
“a	bys” ‘pray’		(*Passio Domini*: line 27)
“a	wra ‘do/will	pysy” pray’	(*Origo Mundi*: line 2197)
“ny ‘not	rug did	cessia” cease’	(Tregear f. 4)
“re	wruk ‘have	scrife” written’	(*Passio Domini*: line 2791)

Figure 138 Particles in the periphrastic verb phrase

The syntactic criteria for identifying determiners are as follows. Most determiners come before adjectival modifiers within the nominal phrase and can be divided into three types: pre-determiners, central-determiners and post-determiners. The demonstrative determiners, -MA and –NA, come after any qualifiers within the nominal phrase. Some other determiners can also come after any qualifiers. Figure 139 shows the syntactic environments in which determiners occur.

modifier				**head**	**qualifier**		
pre-det.	**central-det.**	**post-det.**	**adjective**		**adjective**	**determiner**	
	“un ‘a			den” man’			(*Origo Mundi*: line 94)
	“an ‘the			den man		ma” this’	(*Passio Domini*: line 1306)
	“an ‘the			den man	benegas blessed	ma” this’	(Tregear f.8a)
	“an ‘the	keth same		den man		na” these’	(*Resurrexio Domini*: line 2479)
“ol ‘all	ow my			tus” men’			(*Passio Domini*: line 768)
	“pub ‘each			den man		ol” all’	(*Passio Domini*: line 780)
	“an ‘the		brasse greater	den” man’			(*Passio Domini*: line 773)

Figure 139 Syntactic environments in which determiners occur

Coordinating conjunctions link units of equal grammatical status: word with word, phrase with phrase, or clause with clause. Figure 140 shows the syntactic environments in which coordinating conjunctions occur.

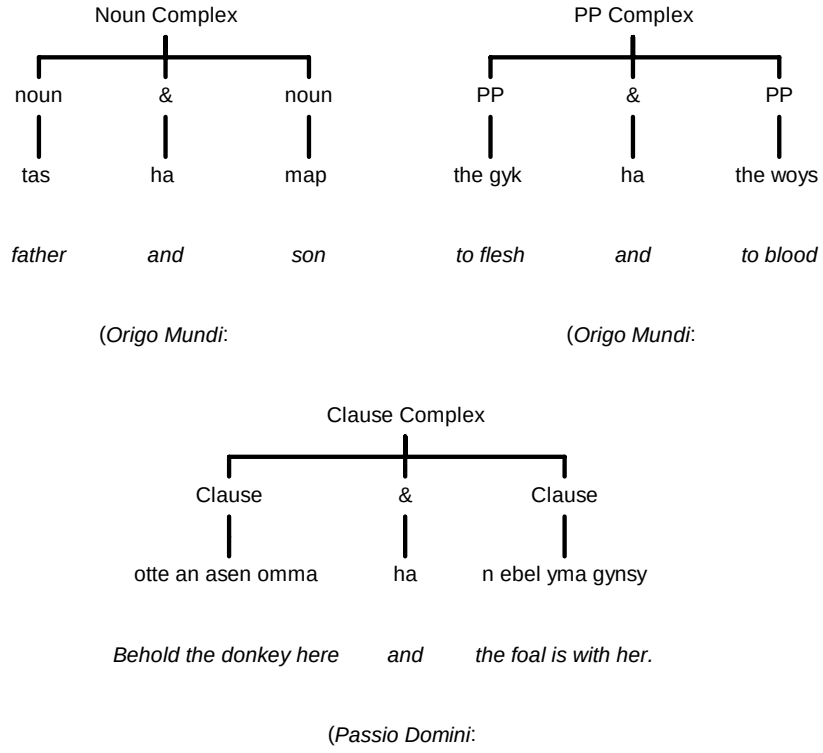


Figure 140 Syntactic environments in which coordinating conjunctions occur

From a syntactic perspective, some items are attested as adjectives, nouns and adverbs. For example, in the following phrase, DA is an adjective.

> "un floch **da"** (*Origo Mundi*: line 664)
>
> 'a **good** child'

But in the following phrase, DA is a noun.

> "mara kyll ze worth an **da** ze wezyll drok agan dry" (*Pascon Agan Arluth*: stanza 21)
>
> 'if he can bring us from the **good** to do evil'

And in the following phrase, DA is an adverb.

“da y won” (*Pascon Agan Arluth*: stanza v.104)

‘well I know’

In the following sentence, CLAF is an adjective.

“Ow colon reseth yn **claf”** (*Passio Domini*: line 1027)

‘My heart is gone **sick’**

But in the following phrase, CLAF is a noun.

“na **claff** vyth ow crowethe” (*Pascon Agan Arluth*: stanza v. 25)

‘nor any bedridden **invalid’**

The base form of the verb can also be used as a noun. Thus in the following phrase, GALLOS is a verb.

“the gorf ker **galles** handle” (*Passio Domini*: line 3194)

‘to be able to hold Thy dear body’

But in the following sentence, GALLOS is a noun.

“scon y **gallos** a vyth lehys” (*Passio Domini*: line 21)

‘soon his **power** will be diminished’

In the following phrase, FYSTENE is a verb.

“ha **fystene** gans touth bras” (*Passio Domini*: line 660)

‘and **hasten** with great speed’

But in the following phrase, FYSTENE is a noun.

“yn un **fystene”** (*Pascon Agan Arluth*: stanza 158)

‘in a hurry’

The past participle may be used as a verb or an adjective. Thus in the

following phrase, *leverys* is a verb.

> “en benenas yn delma yn treze a **leverys**” (*Pascon Agan Arluth*: stanza 253)
>
> ‘so the women **said** to each other’

But in the following phrase, *leverys* is an adjective.

> “zen beth **leverys**” (*Pascon Agan Arluth*: stanza 252) [Adjective]
>
> ‘to the **said** grave’

A program was written to disambiguate certain homographs in the Corpus of Cornish. Syntactic criteria that distinguish part-of-speech may be used to disambiguate certain homographs. A case in point are the homographs *a* and *y*. The verbal particles A and Y only occur preceding a verb, as, for example, “my a vynn” – ‘I want’ (*Beunans Meryasek*: line 12) and “y karsen” – ‘I would like’ (*Beunans Meryasek*: line 15). Sometimes, however, the verb and its particle are separated by a pronoun, as, for example, “my a’n kyv” ‘I will find it’ (*Beunans Meryasek*: line 392). These are the only circumstances under which these verbal particles occur. These facts can be exploited to help disambiguate between the homographs, A (verbal particle), A (conjunction) and A (preposition), and between the homographs Y (verbal particle) and Y (pronoun). Thus, if *a* or *y* is not either immediately followed by a verb or by a pronoun and a verb, then it is not a verbal particle. This may be expressed in predicate logic as follows:

> *l* = lemma.
>
> *t* = token.

n = pronoun.

p = verbal particle.

v = verb.

$B = < t^1, t^2 \mid$ a bigram>

$T = \{ t^1, t^2, t^3 \mid$ a trigram$\}$.

$L = \{l \mid l$ is one of a set of suggested lemmata for a given token $t\}$.

$P = \{$the part-of-speech for a given lemma, $l\}$.

$\exists(t^1)\ \exists(t^2)\ \exists(t^3)\ \exists(l^1)\ \exists(l^2)\ \exists(l^3)$

$((L(l^1 , t^1) \& (L(l^2 , t^2) \& (L(l^3 , t^3))$

$\& ((B(< t^1 , t^2>) \neg P(v, l^1)) \& (T (\{ t^1, t^2, t^3 \mid\}) \& \neg P(n, l^1) \& \neg P(v, l^3)))$

$\rightarrow \neg P(p, l^1))$

This formula is coded in Prolog as follows:

```
not_VP((Txt,Ind),Lem,NewLem):-
 not_VP_VN((Txt,Ind),Lem,NewLem) , not_VP_PN_VN((Txt,Ind),Lem,NewLem).

not_VP_VN((Txt,Ind),Lem,NewLem):-
 vp(VP),
 member((VP,'VP','VP'),Lem),
 Ind2 is Ind+1,
 token_lemmata((Txt,Ind2),Lem2),
 setof(POS,H^D^member((H,D,POS),Lem2),POSs),
 not(member('VN',POSs)),
                                  remove((VP,'VP','VP'),Lem,NewLem).
```

```
not_VP_PN_VN((Txt,Ind),Lem,NewLem):-
 vp(VP),
 member((VP,'VP','VP'),Lem),
 Ind2 is Ind+1,
 token_lemmata((Txt,Ind2),Lem2),
 setof(POS,H^D^member((H,D,POS),Lem2),POSs),
 not(member('PN',POSs)),
 Ind3 is Ind+2,
 token_lemmata((Txt,Ind3),Lem3),
 setof(POS3,H3^D3^member((H3,D3,POS3),Lem3),POSs3),
 not(member('VN',POSs3)),
 remove((VP,'VP','VP'),Lem,NewLem).

vp(a).
vp(y).
```

This short program was found to be an efficient way to speed up homograph disambiguation throughout the corpus. However writing programs of this kind is very time consuming and, therefore, only worthwhile for very frequently occurring items.

5.9 Interlingual Lemmatisation

This section describes the implementation of a Prolog system, *Screffva*, that employs a parallel corpus for the automatic generation of bilingual dictionary entries. The corpus was converted to a Prolog text database and lemmatised. Translation equivalents were then aligned. Finally Prolog predicates were defined for the retrieval of glosses, part-of-speech and example sentences to illustrate usage. Lexemes, including discontinuous multi-word lexemes, are uniquely identified by the system and indexed to their respective segments of the corpus. Glosses and examples of usage can be readily found from the corpus. The system provides a much more powerful research tool than some existing parallel text concordancers.

There are a number of systems available which are capable of producing concordances from parallel corpora. Examples of such systems include

ParaConc, *Wordsmith Tools*, *XCorpus* and *Multiconcord*. With such systems, lemmatisation is usually achieved by entering all the variant forms for the lexeme under investigation. However this method achieves only partial lemmatisation as it is also necessary to disambiguate homographs. Wild cards are often employed to implement a fuzzy search for a particular lexeme; for example **tak*** will find **take**, **takes**, **taking** and **taken**, but not **took**. Use of wildcards also frequently finds words which are not part of the paradigm under investigation. Lemmatisation methods such as these are not precise and post-editing of concordances is, therefore, usually necessary in order to remove unwanted items that are not part of the paradigm being investigated.

Text alignment is usually carried out at the rank of sentence, so that, for example, sentence nine in the first text is equivalent to sentence nine in the parallel text. However alignment at sentence level may prove to be problematic because a translator may translate one sentence by two or more sentences. Alternatively texts can be aligned at the rank of paragraph, so that paragraph nine in one language is equivalent to paragraph nine in its parallel text. Whether alignment takes place at the rank of sentence or paragraph, concordances do not specifically identify translation equivalents of the particular lexemes under investigation. Instead entire sentences which provide the context for the translation equivalents are given.

Existing tools for working with parallel corpora seem, then, to lack two things, firstly the means to uniquely identify lexemes, and secondly the means to identify those items which share translation equivalence in a bitext. These

problems are interrelated since it is between lexemes (rather than the graphical words) of a language pair that equivalence tends to exist. In order to solve these problems, firstly the system must be able to find and retrieve any given segment of text; secondly the bitext must be fully lemmatised; and thirdly the bitext must be aligned at the rank of lexical item.

Screffva consists of a number of modules: the tokenisation module, *VOLTA* - the lemmatisation module, the text alignment module, the corpus, and the dictionary. Lexical items are selected from the corpus on which the dictionary is based whilst simultaneously the dictionary provides information concerning the lemmatisation of the corpus. The processes of dictionary lemmatisation and corpus lemmatisation are, therefore, interdependent. An ideal system for corpus lexicography is one in which the corpus database and the dictionary are interactive. The counterpart of a Prolog dictionary is the lemmatised Prolog text database.

For the purposes of corpus lemmatisation, the lemma should ideally uniquely identify the lexical item. Part-of-speech tagging goes some way towards this. Tagging with the base form further disambiguates the item. Clauses are added to the Prolog text database to represent lemmatisation. *VOLTA* is a program that lemmatises a corpus by relating word tokens as they are encountered in the corpus to a lexicon that has capacity to expand as new items are added. Figure 141 shows the lemmatisation of a short extract from *William Bodinar's Letter*.

```
lemma( (35, 36), ('I',            pron) ).
lemma( (36, 37), (be,             v)    ).
lemma( (37, 38), (a,              art)  ).
lemma( (38, 39), (poor,           adj)  ).
lemma( (39, 40), (fisherman,      n)    ).
```

Figure 141 Lemmatisation of extract from William Bodinar's Letter

Within the bitext, the English word FISHERMAN is translated by the Cornish multi-word lexeme DEN AN PUSKES (literally DEN = 'a man', AN = 'of the', PUSKES = 'fishes'). Figure 142 shows the tokenisation and lemmatisation of the Cornish translation of the English phrase given in William Bodinar's Letter (*William Bodinar's Letter*). The system is very flexible. Not only can the multi-word lexeme, DEN AN PUSKES, be lemmatised as a single lexical item but its component lexemes DEN, AN and PYSK may be simultaneously individually lemmatised. The appearance of the lexeme, PYSK, in its plural form, *puskes*, within the multi-word lexeme, DEN AN PUSKES, does not present a problem to the system. Nor does it present a problem that, due to interruption by the lexeme, BOGHOSEK, the multi-word lexeme, DEN AN PUSKES, is, in this instance, discontinuous.

```
type( (29, 30), thearra ).
type( (30, 31), vee     ).
type( (31, 32), dean    ).
type( (32, 33), bodjack ).
type( (33, 34), an      ).
type( (34, 35), puscas  ).
```

```
lemma( (29, 30), (bones,           v)    ).
lemma( (30, 31), (vy,              pron) ).
lemma( (31, 32), (den,             n)    ).
lemma( (32, 33), (boghosek,        adj)  ).
lemma( (33, 34), (an,              art)  ).
lemma( (34, 35), (pysk,            n)    ).
lemma( (31, 35), ('den an puskes', n)    ).
```

Figure 142 Tokenisation and lemmatisation of translation

The lemma database thus records that the segment between critical points 31 and 35 contains five lexemes: DEN between points 31 and 32, BOGHOSEK between points 32 and 33, AN between points 33 and 34, PYSK between points 34 and 35, and the multi-word lexeme DEN AN PUSKES between

points 31 and 35.

Equivalence exists between the lexical units of a bitext rather than its word types. For this reason alignment takes place between lexical tokens rather than graphic word tokens. Figure 143 shows how equivalents may be entered into the system as 2 place predicates in which the first argument specifies the critical points that bound the Cornish lexical unit, whilst the second argument specifies the critical points that bound the lexical unit which is its English translation.

```
equivalent( (29,30), (36,37) ).
equivalent( (30,31), (35,36) ).
equivalent( (31,35), (39,40) ).
```

Figure 143 Alignment of translation equivalents

Thus the Cornish lexeme, BONES, found at token (29, 30), is translated by the English lexeme, BE, found at token (36, 37). Similarly the Cornish lexeme, VY, found at token (30, 31), is translated by the English lexeme, I, found at token (35, 36). And the Cornish multi-word lexeme, DEN AN PUSKES, found at token (31, 35), is translated by the English single lexeme, FISHERMAN, found at token (39, 40).

It is important to note that this alignment refers to the lexemes listed in the lemma database and does not refer to the types listed in the original tokenisation. If it did, then the English type, *fisherman*, found at token (39, 40), would be the translation of the Cornish phrase, “dean bodjack an puscas”, found at token (31, 35), which is not the case. “Dean bodjack an puscas” translates into English as ‘a poor fisherman’.

A number of predicates are defined which enable the corpus to be used like a dictionary. For example, the predicate, *gloss_lemma/2*, finds the English gloss for a Cornish item or vice versa. The predicate, *entry/1*, displays the full dictionary entry for the requested item (see Figure 144).

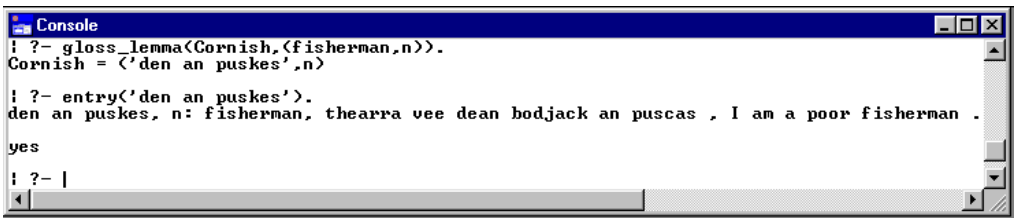


Figure 144 Using the *Screffva* system

The *Screffva* system is able to retrieve any given segment of text, and uniquely identifies lexemes and the equivalences that exist between the lexical items in a bitext. Furthermore the system is able to cope with discontinuous multi-word lexemes. The system is thus able to find glosses for individual lexical items or to produce longer lexical entries which include part-of-speech, glosses and example sentences from the corpus. Insofar as the system is able to identify specific translation equivalents in the bitext, the system provides a much more powerful research tool than existing concordancers such as *ParaConc*, *WordSmith*, *XCorpus* and *Multiconcord*. The system is able to automatically generate a bilingual dictionary which can be exported and used as the basis for a paper dictionary. Alternatively the system can be used directly as an electronic bilingual dictionary.

6 Conclusion

Lemmatisation entails firstly a prior knowledge of the inflectional system necessary for base form lemmatisation, and secondly a prior knowledge of syntax, morphology, phonology and lexical semantics necessary for the identification of part-of-speech. This grammatical knowledge needs to account for every lexical item in the corpus. Existing grammars may prove insufficient for the task in hand, as was the case with the corpus of Cornish. When existing grammars prove to be insufficient, then some sort of bootstrapping is necessary. Even for a language which has been well described, such as English, a corpus may contain nonce forms and usages, which are not described in existing grammars. So this bootstrapping problem is likely to be found to a greater or lesser extent in all lexicographical work.

No fully automatic corpus lemmatisation system has been developed which is 100% accurate. The results of fully automatic systems, therefore, need to be checked by human lexicographers. There is much to be said, then, for a system which involves human interaction during the lemmatisation process, since checking takes place simultaneously with the lemmatisation process.

If we are to understand the process of lemmatisation in relation to the Cornish language, it is necessary first to look at the traditions of Cornish lexicography. The history of Cornish lexicography places lemmatisation in a social as well as cognitive perspective. Cornish lexicographical practice has evolved to develop social norms as well as to provide lexical explication. The circumstances in which Cornish lexicography has taken place have a

bearing on the manner in which Cornish lexicography has been undertaken. During the eighteenth century, Cornish lexicography was to a large extent undertaken within an environment of Cornish antiquarian scholarship. In the late nineteenth century, the broader backdrop of Celtic studies was the setting for Cornish lexical investigation. Then in the twentieth century, Cornish language revival became the driving force for lexicographical activity. Throughout history, Cornish lexicography has been focussed on translation; even onomastic Cornish dictionaries concentrate on attempting to translate names into English.

Examples of Cornish lexicography are only extant in the Old Cornish period, the Modern Cornish period and the 20th century. Cornish lexicography has its beginnings in the marginal Old Cornish glosses found in Latin documents between the ninth and eleventh centuries. The next stage in the evolution of Cornish lexicography is the bringing together of such glosses to form a glossary. The *Vocabularium Cornicum* (VC) of circa 1100 AD is an example of such a glossary. Presumably the practice of making and using glossaries continued through the Middle Cornish period. However no extant glossaries from the Middle Cornish period have been discovered. A few examples of non-alphabetical glossaries are found in the 17th century (British Library Add. 17062; National Library of Wales, Bodewryd MS5).

A major event in the development of Cornish lexicography is the translation of the Middle Cornish corpus by John Keigwin at the end of the seventeenth century. Keigwin translated *Pascon Agan Arluth* at the request of Sir Francis

North in 1678. In 1693, Keigwin translated Jordan's *Gwreans and Bys* and the *Ordinalia* at the request of Bishop Jonathan Trelawney. It is these translations of Keigwin's that form the core of the corpus used by Lhuyd for the Cornish in his *Archaeologia Britannica* (AB). Lhuyd, however, does not give Cornish items in their attested spellings, but respells them in his own phonetic notation. Tonkin then used Lhuyd's *Archaeologia Britannica* (AB) as the main source for his Cornish-Latin-English vocabulary (CLEV). Tonkin does not use Lhuyd's phonetic notation, but spells Cornish items according to the Modern Cornish orthographic conventions of his own time. Pryce (ACB) used Tonkin's vocabulary as the basis for his "Cornish-English Vocabulary". Norris (1859a) used Pryce's "Cornish-English Vocabulary" (ACB) and Keigwin's translation to produce his own edition and translation of the *Ordinalia*. Stokes also used Pryce's vocabulary (ACB) and Keigwin's translations to produce his own editions and translations of *Pascon Agan Arluth* (Stokes 1861) and *Gwreans an Bys* (Stokes 1863). Robert Williams used the published editions by Norris (1859a) and Stokes (1861; 1863) together with Lhuyd's *Archaeologia Britannica* (AB) and Tonkins's vocabulary (CLEV) as the corpus for his *Lexicon Cornu-Britannicum* (LCB). Jago then reversed Williams' *Lexicon Cornu-Britannicum* (LCB) to create his *English-Cornish Dictionary* (ECD1). In *An English Cornish Dictionary* (ECD2), Morton Nance and Smith (ECD2) put the Cornish equivalents in Jago's *English-Cornish Dictionary* (ECD1) into normalised orthography using Morton Nance's (1929) Unified Cornish spelling system. Morton Nance then reversed the ECD2 to create his *New Cornish English Dictionary* (NCED). In his *Gerlyver*

Kernewek Kemmyn (GKK), George puts the head words of Morton Nance's NCED into George's own Kernewek Kemmyn spelling system. Thus it can be seen how dictionaries down to the present day are largely derived from a common source, the corpus of translations made by John Keigwin at the end of the seventeenth century.

Of course, Keigwin's translations are not the only source used by Cornish lexicographers. Lhuyd used other sources apart from Keigwin, including the Rev. Henry Ustick, James Jenkins and Nicholas Boson. Lhuyd also incorporated the Old Cornish *Vocabularium Cornicum* (VC) into his corpus. Various fragments of Modern Cornish provide additional sources for the vocabularies of Tonkin (CLEV), Gwavas (*Gwavas Manuscripts*: 119v to 125r) and Borlase (VCBL). Twentieth century Cornish dictionaries have benefited from the discovery of *Beunans Meriasek* and the *Tregear Homilies* to expand their corpora. In addition twentieth century lexicographers have profited from articles published in various scholarly journals from the end of the nineteenth century onwards. The twentieth century also saw a steady increase in the inclusion of neologisms.

Since the beginning of the eighteenth century, fascination with Cornish place names and personal names has caused a steady stream of onomastic dictionaries to be produced. These dictionaries are mainly concerned with giving the etymologies of place names and personal names. The etymologies given in Cornish onomastic dictionaries are largely conjectural and are based on assonance. The compilers of Cornish onomastic dictionaries tend not to

distinguish between the etymology of an onomastic term and its meaning. Onomastic terms are referring expressions; they denote people or places. Thus the sentence, “Peter visited Bridgend” means that Peter visited a particular geographical location that is named “Bridgend”. It does not mean that Peter visited the end of a bridge. However Cornish lexicographers frequently refer to onomastic etymologies as “meanings” (GCN: vi *et passim*; GCPN: 1 *et passim*; TCPNE: v; PDCPN: 47 *et passim*; FNWP: 9 *et passim*; CPNL: 191), “interpretations” (CN: 9; GCPN: 1; FNWP: 15), or “significations” (Gwavas 1738; *Mems. of the Cornish Tongue*: Part II, 128; GCN: viii). In the twentieth century it came to be felt that systematic analysis of Cornish names is hampered by the capricious spelling of attestations. Some Cornish lexicographers (GCPN; CPNE) attempt to circumvent this problem by giving onomastic terms only in normalised orthography. Others (TCPNE; PNWP, FNWP) gloss the attested form with its equivalent form in normalised spelling. Most Cornish onomastic dictionaries do not give information regarding pronunciation. This seems to be a serious omission since the pronunciation of Cornish place names is frequently far from obvious from their written form.

The texts that have been included in the Corpus of Cornish are from the Middle Cornish and Modern Cornish periods. They cover a period ranging from the late 14th century to the latter part of the 18th century with, in addition, a couple of tiny fragments from the 19th century. It is important to understand the nature and characteristics of these texts that comprise the corpus. Much of the variation in orthographic practice is attributable to the diachronic range that the corpus encompasses. In addition to this diachronic

variation, one finds considerable evidence of capricious spelling even within a single document. Published critical editions of the source texts are of varying reliability and it was deemed necessary to create new critical editions for the electronic corpus. In particular, these new critical editions incorporate a new lexicon based tokenisation.

The number of informants represented in the corpus is small. This is especially true of the Middle Cornish period. Of the six texts that comprise the Middle Cornish component of the corpus, only one, *Beunans Meriasek*, bears a colophon to identify its author. It is not clear to what extent each individual Middle Cornish text is the work of a single author. Nor is anything much known about the authors of the Middle Cornish texts; one cannot even be certain that they were mother-tongue speakers of Cornish. The Modern Cornish component of the corpus, on the other hand, has a far greater number of informants. Biographical details are known for several of the Modern Cornish informants, including whether or not they were mother-tongue speakers of Cornish. It might be felt that some informants, those known to be mother tongue speakers of Cornish, for example, are more reliable than others. This might influence a lexicographer in the choice of canonical form from several attested base forms.

Due to the scarcity of source Cornish texts, it is considered desirable to include all the extant material available. This inevitably leads to a corpus that is quantitatively unbalanced with regard to diatextual features and with regard to diachronic representation. The Middle Cornish component of the corpus

contains a very limited range of genres. A relatively unexplored area is that of the diastratic and diaphasic information that might be derived from the corpus. The miracle plays in particular contain parts for kings, courtiers, members of the aristocracy, servants, artisans, God, Christ, angels, saints, bishops, and other members of society. This fact may make it possible to ascribe certain members of synonym sets to the acrolectal or basilectal end of the diastratic continuum or to particular registers.

New critical editions of the source texts were created for the Corpus of Cornish. The methodology of tokenisation proved to be a central issue in the compilation of the corpus. The disparity between orthographic words and morphosyntactic words drew attention to the necessity of a systemic method of tokenisation for lexicographical purposes. The notions of token, type and tone provide a framework for rendering a handwritten manuscript into an electronic tokenised critical edition. For lexicographical purposes, the unit of tokenisation is the lexical item. Lexical items appear on a scale of rank as either morphemes, words or multi-word lexemes. The identification of multi-word lexemes for tokenisation purposes is not trivial. Multi-word lexemes fall into two categories, collocations and fixed expressions and a number of criteria are suggested for their identification.

Two algorithms for corpus tokenisation were considered, character based tokenisation and lexicon based tokenisation. It was noted that, based as it is on the orthographic word, character based tokenisation does not cope well with the three ranks at which lexical items occur. Lexicon based tokenisation, on

the other hand, is able to cope with items that are realised at different points on the scale of rank. Furthermore lexicon based tokenisation is able to identify instances of combinatorial and overlapping ambiguity. Tokenised texts are represented by means of critical tokenisation as Prolog files. The resulting text database can be manipulated in Prolog to retrieve types and tokens, and to make word lists and concordances.

The lemma has been considered from two perspectives. Firstly it has been considered according to the available literature on lexicographical theory. And secondly it has been considered according to the history of Cornish lexicography. One of the main concerns of lemmatisation is to bring the variant forms of the lexeme together under a single canonical form. Cornish lexical items are attested in a wide variety of forms. This lexical variation is either synchronic or diachronic.

Synchronic variation is concerned with inflection and derivation as well as conditioned variation and free variation. In Cornish, countable nouns, verbs, prepositions, adjectives and cardinal numbers may be inflected. Cornish inflection may be realised by a prefix, a suffix, an infix, a suprafix or by vowel affection. It should be explicit from the lemma to what declension or conjugation the item belongs. Any irregularities or suppletion in the paradigm should also be indicated in the lemma. Conditioned variation includes initial mutation of consonants and for some items apocope. Free variation is alternation in form which is not in any way systemic. Free variation of the base form is particularly troublesome for the lexicographer who will have to

decide whether to give variant spellings of the base form separate entries or to cross reference them to a canonical form. The boundary between inflection and derivation may be unclear and tradition may be the deciding factor. Derivatives may be given full entries, may be included in the entry for the main form, or may be omitted from the word list.

Diachronic variation refers to differences in form between etyma attested over a period of time. Metathesis, intrusion, elision and mutation comprise the main elements in a typology of Cornish diachronic variation. Metathesis involves the transposition of syllables or phonemes. Intrusion involves the insertion of a segment and may be of three types. Prothesis involves the insertion of a segment at the beginning of a word; epenthesis involves the insertion of a segment into the middle of a word; and paragoge involves the insertion of a segment at the end of a word. Elision involves the omission of a segment from a word. Loss of an initial segment is known as aphesis; loss of a medial segment is known as syncope; and loss of a final segment is known as apocope. Two types of diachronic mutation are frequently attested in Cornish. Between the Middle and Modern Cornish periods we often encounter diphthongisation, the replacement of a simple vowel by a diphthong. Secondly Middle Cornish <M> and <N> are often found pre-occluded in Modern Cornish. This framework of metathesis, intrusion, elision and mutation provides a system for determining whether items are etymologically related.

The entry form is the form which begins a dictionary entry and determines that entry's place in the word list. An entry form may be either a base form or an

oblique form. The base form is the part of the paradigm that is chosen to represent the lexeme. Tradition usually determines which part of the paradigm serves as the base form. Ideally the part of the paradigm chosen for the base form should include some indication of the declension or conjugation. Of the various attested spellings of the base form, the canonical form is the one preferred or chosen by the lexicographer. Choice of canonical form may be based on either prescriptive or normative principles. Posited, authoritative norms form the basis of the prescriptive principle. The normative principle, on the other hand, draws on regular usage as attested in a corpus to establish norms.

Whilst an alphabetically sorted word list does not represent the semantic relations between lexical items, alphabetical order is considered to be the fastest and easiest system for the dictionary user. Within an alphabetically arranged word list, irregular and suppletive oblique forms need entries within the word list which cross-reference to their canonical form.

The lexicographer has a choice of either giving a derived form a main entry in the alphabetically arranged word list or listing the derived form as a run on. Nests are created by regrouping derivatives under the head of the word family. Nests of this kind save space. A difficulty arises when very extensive use is made of rich and large nests since such nests can disagree quite significantly with the alphabetical arrangement.

Concerning compounds and multi-word lexemes, the lexicographer must decide whether a group of words is sufficiently stabilised to treat as a single

lexical item. There are three methods for locating compounds and multi-word lexemes in the word list. Firstly, the compound or multi-word lexeme may be treated in the same way as any other string of letters. Secondly, such items may grouped together as a block by listing them after their first word. Thirdly they may be listed under the element which is considered to be the most important. Alternatively compounds and multi-word lexemes may be treated as sub-entries; this facilitates alphabetical insertion by the second or third word.

The lemma in Cornish glossaries and dictionaries evolved to become increasingly more systematic between the 18th and the 20th centuries. The earlier glossaries of Cornish are not sorted alphabetically at all. From the eighteenth century onwards, we find Cornish glossaries sorted alphabetically by the first letter of the entry word or by the first two letters for larger glossaries. Full alphabetical sorting of the word list appears with the advent of the first published glossaries in the latter half of the eighteenth century.

During the course of the eighteenth and nineteenth centuries the alphabet itself became redefined. In the eighteenth century, the practice of treating <I> and <J>, and <U> and <V> as homographic starts to change. We find <I> and <U> being used where they are presumed to be vocalic, and <J> and <V> are used where they are presumed to be consonantal. Eighteenth century Cornish lexicographers continue, however, to conflate <I> and <J>, and <U> and <V> for the purposes of sorting the word list. In the nineteenth century, this conflation was finally abandoned and initial <I>, <J>, <U> and <V> each

have their place in the macrostructure. Thus the criterion of putative pronunciation marks the first phase in the development of a standardised orthography. Williams (LCB) makes further adjustments to Cornish orthography by respelling <K> as <C>, and respelling <ʒ> as either <TH> or <DH> throughout.

The first completely normalised orthography is Morton Nance's (1929) Unified Cornish. Unified Cornish retains Williams' <TH> and <DH> but reintroduces <K>. In Unified Cornish, <C> before <A>, <O>, <U>, <L> or <R> is pronounced /k/, but before <E> or <I>, <C> is pronounced /s/. Thus <K> is necessary before <E> or <I> to indicate when the pronunciation is assumed to be /k/. This heterographic distribution of <C> and <K> is consistent with Middle Cornish orthographic practice. George's (1984, 1986) Kernewek Kemmyn orthography is based on his conjectural Middle Cornish phonology. In the case of Kernewek Kemmyn, respelling is far more extensive than is found in Unified Cornish. Gendall's (PDMC, NPDMC) normalised orthography is achieved by choosing a canonical form from amongst the base forms attested in Modern Cornish. Gendall avoids respelling wherever possible.

In the 18th and 19th centuries, both base and oblique forms, and both mutated and radical forms comprise the head word list; the semantic unit represented by the head word may be a vocable, a lexeme or a lexical unit; and on the scale of rank, the head word may be a multi-word lexeme, a word or a morpheme. In the 20th century, head words are generally given in the base

form; oblique forms tend to only appear as head words when they are not attested in the base form; irregular oblique forms are usually cross-referenced to their canonical form. It is only in the 20th century that appendices, containing tables of mutations, and paradigms of verbs, pronouns and prepositions appear in Cornish dictionaries .

In the earliest Cornish glossaries, the lemma consists of the head word only. Thereafter lexicographers of Cornish start to add more fields to the lemma. In the eighteenth century we start to find variant base forms given after the canonical form. Williams (LCB) is the first lexicographer of Cornish to include a part-of-speech field in the lemma. Morton Nance (NCED) also includes part-of-speech and introduces several more fields: etymology, oblique forms and mutation numbers. In point of fact the etymological information given by Morton Nance is found in two fields. The first, placed before the head word, uses the symbols † and * to indicate respelling from Old Cornish or neologism borrowed from Welsh or Breton. The second etymological field follows the head word and contains the etymon and its source. Morton Nance does not include a separate field for pronunciation but employs diacritics over the head word for this purpose. George (GKK) introduces three more fields: a homograph disambiguator; an authentication code; and for a few entries only, phonetic transcription. George's homograph disambiguator usually consists of either an English translation equivalent or the part-of-speech, and, since this information only duplicates what is included elsewhere in the entry, this disambiguator serves no useful purpose. George's authentication code includes etymological information concerning the

frequency of attestation and sources. Gendall (PDMC) is the first to include phonetic transcription in all lemmata. Gendall also distinguishes homographs by allocating a number to each of them. Part-of-speech, phonetic transcription, oblique forms, variant base forms, mutation numbers and etymological information may all serve to distinguish between homographs. In the few cases where these fields prove insufficient, a genre field label could be included to provide final disambiguation between homographs.

System networks have been used in this project to fulfil several functions. The unit of lemmatisation system governs tokenisation of the corpus and the ranks at which items are included in the word list. The entry form system governs which grammatical forms of the lexical item may be used as entry forms. The derivative entry system governs the manner in which derivatives are entered in the macrostructure. The synchronic variation system network is very large being comprised of the Cornish inflection system network, the synchronic mutational variation system network and the apocope system. The synchronic variation system network generates the base form from a given word type. The Cornish inflection system also contributes to the generation of the part-of-speech field in the lemma; though part-of-speech is also determined by a range of other criteria including syntax, semantics and phonology. On the one hand, these system networks are descriptive of what is found in Cornish dictionaries. On the other hand they are generative, providing the means to generate the dictionary macrostructure from the corpus. System networks of this kind are largely based on the Boolean operators, AND and OR. Traversing a system network is essentially algorithmic and thus system networks may be readily

converted to provide computer programs.

Although system networks are generative, full automation of base form lemmatisation was not possible for the corpus of Cornish since all the spelling variants of morphemes were not known until lemmatisation was complete. System networks do, however, provide the human lexicographer with a grammatical framework for classifying words under their lexemes.

Part-of-speech tagging goes some way towards separating homographs. However, if the purpose is to uniquely identify all the lexemes attested in the corpus, then a more thorough system of lemmatisation is required. When inserting lexeme tags directly into text, issues of segmentation arise. Contractions need to be decomposed and multi-word lexemes need to be treated as wholes. Punctuation marks may be treated as separate words.

Dictionaries frequently include forms other than the preferred/canonical base form in the word list. Such practice may result in multiple entries for a single lexeme being scattered throughout the word list. Depending on a dictionary's intended use, this practice may be perfectly satisfactory. If, however, one wants to provide a unique code for every lexeme attested in a corpus, in order to tag that corpus, then a word list that includes multiple entries for a single lexeme is unsatisfactory. Lemmata for corpus lemmatisation thus need to consist of a canonical base form followed by homograph disambiguators, such as part-of-speech and semantic-field. If a word list from an already existing dictionary is to be used for lemmatising a corpus, then it will possibly have to be adapted before lemmatisation begins, so that it does not contain multiple

entries for any single lexeme. Furthermore, such a word list will have to be extended during the lemmatisation process, to include items attested in the corpus that were not included in the original dictionary from which the word list was obtained.

There are essentially two approaches to the lemmatisation of computerised text corpora. With the first approach, a computerised dictionary is used as a lemmatisation database that relates word types attested in the text corpus to their base forms or lemmata. The Prolog programming language provides a suitable tool for the construction of such a database. Manual disambiguation of homographs is necessary with this method. The second approach employs morphological rules that govern the combinatorial association of free morphemes and affixes. Two problems are typically encountered. The first problem concerns contractions resulting from elision or assimilation. The lexemes that comprise these contractions need to be separated. The second problem concerns the disambiguation of homographs. Contexts supply the relevant data for disambiguation of homographs.

In order to achieve full base form lemmatisation, first all the variant forms of a lexeme must be identified; secondly homographs need to be distinguished. Three methods were trialed, all of which are capable of producing a lemmatised corpus.

The first method that was devised for this project, the *VOLTA* system, employs a lemmatisation database that expands as lemmatisation takes place and new items are encountered. A word token is input and this is checked against the

database. If a word type is found in the database a list of possible lemmata are offered and one is chosen. If the word type is not found in the database, then the operator supplies the lemma, which is written to the database. The word type and the lemma are then output to a text file to create the lemmatised version of the text. *VOLTA* is not fully automatic. However, it has the advantage that the lexicon is bootstrapped from the corpus as the lemmatisation proceeds. With this method, a separate look-up-lexicon needs to be generated for each text since there is a great deal of difference in spelling conventions between the individual texts that comprise the corpus, especially when texts are from different historical periods.

Method two involves use of normalised spelling. Inconsistent spelling causes a problem for the corpus linguist. This is especially the case with the Corpus of Cornish. Putting the corpus into a normalised spelling system makes it very much more accessible. In normalised orthography, the Corpus of Cornish contains far fewer word types and also far fewer homographs than in its original orthography. The corpus in normalised spelling can be aligned with the corpus in its original orthography, thus enabling access of the original version via normalised spelling. This second method has the advantage that a single look-up-lexicon may be used for the entire corpus. Once the corpus in its original orthography has been aligned with the corpus in normalised spelling, the process of lemmatisation is much faster than the first method. This is for two reasons. Firstly, since each lexeme has fewer variant forms when the corpus is normalised spelling, the resulting lexical database is correspondingly smaller. Secondly, there are far fewer homographs that need

to be distinguished manually by the lexicographer.

The third method has the disadvantage that a great deal of work must be undertaken to compile the morphological rules before lemmatisation can take place. Furthermore, this method only achieves a partial lemmatisation since it does not distinguish homographs.

With the text in normalised orthography, the application of morphological rules can be used to achieve partial lemmatisation. This method only works where there is regularity in the inflexional paradigm. In the case of Cornish, rules governing the mutation of initial consonants have to be taken into account as well as inflectional affixes. Homographs also have to be distinguished. The most efficient way to apply morphological rules is to first generate a list of all the word types in the corpus; then apply the rules to each word type to create a look-up dictionary that relates types to their base forms. The efficiency of this morphological word type parser is much improved if the lemmata that are generated by the system are checked against a dictionary of base forms and their part-of-speech. Due to the homography of affixes, however, human intervention is still necessary to correct errors that are generated by the system.

Two approaches to the creation of a morphological parser were tried and compared. The stochastic approach, employing *Linguistica* software, was found to be not very reliable. The morphological database generated with *Linguistica* accounts for less than half of the word types attested in the corpus. Furthermore considerable human intervention would be necessary if the

database generated with the help of *Linguistica* were to be used to lemmatise the corpus. The manually created morphological parser was found to be far better. It left fewer word types for which no lemma is suggested and fewer types requiring disambiguation as a result of more than one lemma being suggested. Although the manually created morphological parser might be improved by the addition of more rules, no parser based on affix stripping can handle suppletion or purely capricious irregularity. A morphological parser is thus only a partial solution to the creation of a lemmatisation database.

Of the three lemmatisation methods that were trialed, the third was the least successful since lemmatisation is only partial. Of methods one and two, there is a trade off between additional time needed to align the corpus in its original orthography with the corpus in normalised spelling, and time saved by applying the lemmatisation process to the normalised corpus.

Semantic distinctiveness, etymology, and grammatical difference may all be used to distinguish between homographs. Semantic distinctiveness is determined by mother-tongue speakers of a given language being unable to perceive any relationship between different senses. Only pairs with extremely diverse unconnected senses should be recognized as homographs. In the case of the historical Cornish that comprises the corpus, there are no first language speakers today who could act as informants concerning semantic distinctiveness. Etymology provides another criterion for distinguishing homographs. However etymology is frequently uncertain or merely conjectural.

For the corpus of historical Cornish, grammatical difference is thus the most reliable way of separating homographs. Grammatical differences include difference in part-of-speech, or, if the part-of-speech is identical for a pair of homographs, a difference of conjugation or declension. Determining part-of-speech is, however, not a trivial matter. Semantic, phonological, morphological and syntactic criteria may all be used to identify part-of-speech. Though these criteria can produce different results. Morphological criteria are preferred for lexicographical purposes, since the lemma represents an inflectional paradigm. For items which do not inflect, however, syntactic or other criteria may be employed.

It was found that the disambiguation of at least some homographs may be reliably automated by the use of computer programming. Writing computer programs of this kind is, however, time consuming and, therefore, only worthwhile for frequently occurring items. For less frequently occurring homographs, manual disambiguation is quicker.

Interlingual lemmatisation refers to the alignment of translation equivalent lexemes in a bitext. The *Screffva* software developed for this project was found to provide an adequate means for storing such an alignment and for accessing translation equivalent lexemes. *Screffva* permits both multi-word lexemes and their component lexemes to be simultaneously lemmatised. Discontinuous lexemes also present no problem to the *Screffva* system. Lemmatisation of both texts in the bitext is essential, since it is lemmata that define the token boundaries of individual lexemes. The *Screffva* system is

capable of the automatic generation of a bilingual dictionary.

Linguistic and lexicographical terminology is structured in such a manner that the way in which we conceptualise is influenced. Terminology evolves through a process of historical accretion. Thus the history of lexicography plays an important part in the way in which today's Cornish lexicographers operate.

A dictionary is both a social and a cognitive artefact. As a social artefact it reflects the norms of a community. These norms were acquired by the community through a process of historical accretion. Alphabetical order is an example of one such norm. In societies with alphabetic writing systems, alphabetical order is learned from a very early age so that members of that society feel comfortable with alphabetically arranged macrostructures. The idea that, for the purposes of lemmatisation, a base form can be used to represent a paradigm is another commonly accepted norm. The move towards normalised orthography and preferred spellings of the base form is yet another example of lexicographical norm. It is noted that, in the case of Cornish, this latter process of normalising orthography is still taking place, since a number of competing orthographies exist for revived Cornish. Cornish lexicography shares many such norms with European lexicography in general and it may well be that these norms have been simply transferred from the lexicography of other languages such as English and Latin. Since the history of Cornish lexicography has been entirely bilingual, the tendency has been for grammatical categories to be transferred from English to Cornish.

Tokenisation of the corpus has also been influenced by English lexical equivalence.

As a cognitive artefact, a dictionary provides lexical explication. A dictionary also embodies much linguistic theory with regard to morphology, part-of-speech, etc.. We have seen how there has been a steady accumulation of linguistic and lexicographical concepts throughout history and these concepts are accompanied by an accretion of terminology. However many of these lexicographical and linguistic concepts are nebulous by nature. Nebulous concepts of this kind include the very notion of the word itself. The distinction between inflection and derivation is vague. The notion of part-of-speech is another imprecise concept. Several criteria may be employed for the purpose of identifying part-of-speech. However these criteria do not always achieve the same result. A verb defined as such by morphological criteria might be an adjective or a noun if syntactic criteria are invoked. Thus we have morphological-verbs and syntactic-verbs, morphological-nouns and syntactic-nouns, and so on. For lexicographical purposes, morphological part-of-speech is particularly relevant when, for lemmatisation purposes, a base form is used to represent a paradigm. Morphological part-of-speech, however, does not distinguish between those word classes which do not inflect. Lexicographers, therefore, invoke other criteria to distinguish non-inflecting word classes. However lexical items are normally classified in dictionaries simply as nouns, verbs, adjectives, etc. without any explanation of which criteria are being used for the purposes of classification. This obscures the fact that a morphological-noun is, strictly speaking, not the same thing as a syntactic noun. The problem

lies in our terminology; and our terminology, in turn, influences our cognition and hinders us from perceiving language as it really is. One might, then, question the value of the part-of-speech field in dictionaries when the underlying concept of lexicographical part-of-speech is so inherently nebulous. It might be argued that ultimately our familiar part-of-speech categories should be abandoned and replaced with new terms for morphological word classes, syntactic word classes, etc.. This would lead to the part-of-speech field in the dictionary lemma being replaced by several new fields: morphological word class, syntactic word class, semantic word class, phonological word class. However such a drastic change would very likely meet with resistance from dictionary users, for whom the old part-of-speech categories are familiar. A proliferation of lemma fields with strange new categories would possibly deter many dictionary users. In order for radically new ideas to be accepted by a community, they first have to undergo peer review at the cognitive level and this is a very slow process. A practice sometimes adopted by lexicographers is to combine a number of part-of-speech categories within a single entry: verb and noun (verbal noun), noun and adjective, adjective and adverb.

The inherent fuzziness of lexico-linguistic concepts makes it difficult or impossible to create a completely reliable computer algorithm. A tabula rasa is first required followed by the conception of logically discrete concepts and terms to refer to these concepts. Many writers use the terms ‘lexeme’, ‘lemma’, ‘head word’, ‘base form’, and ‘canonical form’ as synonyms. I have sought to disambiguate these terms for the purpose of this study in order to

provide a more precise denotation of concept.

In conclusion, the methods and techniques that can be brought to bear on the historical corpus of Cornish to generate a lemmatised dictionary macrostructure are as follows. An investigation of lexicographical history and tradition is a prerequisite since it plays an important part in the way in which Cornish lexicography is practised today. It is lexicographical history and tradition that defines the alphabet that is used, the alphabetical order of the macrostructure, the choice of grammatical form used as the base form, and the fields that constitute the lemma. Tokenisation at the rank of lexical item is the first stage in the process of lemmatisation. Since it copes with the ranks of morpheme, word and multi-word lexeme, lexicon based tokenisation is to be preferred over character based tokenisation. After tokenisation, lemmatisation essentially consists of two operations: the generation of the base form, and the disambiguation of homographs. Regarding the first of these operations, system networks provide the means to generate base forms from attested word types. Regarding the second operation, grammatical difference is the most reliable way of disambiguating homographs. The provision of a unique code for every lexeme attested in the corpus may be accomplished by a lemma that contains three fields: the canonical form, the part-of-speech and a semantic field label. Whilst computer programs are an extremely useful aid to lemmatisation, they are not usually fully automatic with 100% accuracy. Although, in theory at least, it ought to be possible to write a program that would lemmatise a corpus with 100% accuracy, the level of linguistic detail that would need to be incorporated in the program would require that the corpus first be lemmatised

before the program could be written. A solution is to employ programs which involve human interaction during the lemmatisation process, thus allowing the lemmatisation database to be bootstrapped as lemmatisation takes place. Computerised morphological processing may be used at least to partially create the lemmatisation database, especially if the corpus is available in normalised orthography. Disambiguation of at least some of the most common homographs may be automated by the use of computer programs. Interlingual lemmatisation provides the means to generate the macrostructure for both sides of a bilingual Cornish-English and English-Cornish dictionary, and to supply translation equivalents and example sentences for each lemma.

Bibliography

Cited Dictionaries

AAOELG AN EDITION OF ABBOT AELFRIC'S OLD ENGLISH-LATIN GLOSSARY WITH COMMENTARY (1983) ed. R. Gillingham. Ohio State University: PhD thesis; Ann Arbor: UMI (1983).

AB ARCHAEOLOGIA BRITANNICA: VOL. I GLOSSOGRAPHY (1707) comp. Edward Lhuyd. Oxford: The Theatre.

ACB ARCHAEOLOGIA CORNU-BRITANNICA; OR, AN ESSAY TO PRESERVE THE ANCIENT CORNISH LANGUAGE; CONTAINING THE RUDIMENTS OF THAT DIALECT, IN A CORNISH GRAMMAR AND CORNISH-ENGLISH VOCABULARY, COMPILED FROM A VARIETY OF MATERIALS WHICH HAVE BEEN INACCESIBLE TO ALL OTHER AUTHORS WHEREIN THE BRITISH ORIGINAL OF SOME THOUSAND ENGLISH WORDS IN COMMON USE IS DEMONSTRATED; TOGETHER WITH THAT OF THE PROPER NAMES OF MOST TOWNS, PARISHES, VILLAGES, MINES, AND GENTLEMEN'S SEATS AND FAMILIES, IN WALES, CORNWALL, DEVONSHIRE, AND OTHER PARTS OF ENGLAND (1790) comp. W. Pryce. Sherborne: The Author.

ALB ANTIQUAE LINGUAE BRITANNICAE, NUNC VULGÒ DICTAE CAMBRO-BRITANNICAE, A SUIS CYMRAECAE VEL CAMBRICAE, AB ALIIS WALLICAE ET LINGVAE LATINAE DICTIONARIUM DUPLEX. PRIUS, BRITANNICO-LATINUM, PLURIMIS VENERANDÆ ANTIQUITATIS BRITANNICÆ MONUMENTIS RESPERSUM. POSTERIUS LATINO-BRITANNICUM. ACCESSERUNT ADAGIA BRITANNICA, & PLURA & EMENDATIORA, QUÀM ANTEHÀC EDITA (1632) comp. John Davies. Londini: impress. in aedibus R. Young, impensis J. Davies.

ALDC THE ANCIENT LANGUAGE AND DIALECT OF CORNWALL (1882) comp. F. Jago. Truro: Netherton and Worth.

CDS1 CORNISH DICTIONARY SUPPLEMENT NO 1: PYTHOW AND GEGYN (KITCHEN THINGS) AND WAR AN FORDHOW (ON THE ROADS) (1981) comp. J. Snell & W. Morris. N.p.: Cornish Language Board.

CDS2 CORNISH DICTIONARY SUPPLEMENT NO. 2: CHY HA SOTHVA (HOME AND OFFICE) (1984) comp. J. Snell & W. Morris. N.p.: Cornish Language Board.

CDS3 CORNISH DICTIONARY SUPPLEMENT NO. 3: GERYOW DYVERS (1995) comp. W. Morris. N.p.: Agan Tavas.

CED CORNISH-ENGLISH DICTIONARY (1955) comp. Robert Morton Nance. N.p.: Federation of Old Cornwall Societies.

CEV A CORNISH-ENGLISH VOCABULARY; A VOCABULARY OF LOCAL NAMES, CHIEFLY SAXON; AND A PROVINCIAL GLOSSARY comp. R. Polwhele. In R. Polwhele *The History Of Cornwall* VI. (1808) London: Cadell and Davies.

CG1 A CORNISH GLOSSARY (1868) comp. Whitley Stokes. London: Asher.

CG2 A CORNISH GLOSSARY (1869) comp. Whitley Stokes. *Transactions of the Philological Society* 137-250. Oxford: Basil Blackwell.

CLEV A CORNISH-LATIN-ENGLISH VOCABULARY (n.d.) comp. Thomas Tonkin. Bilbao MSS., copy at the Royal Institution of Cornwall (MS. "B").

CN CORNISH NAMES (1926) comp. T.F.G. Dexter. London: Longmans, Green.

CNACD A COLLATION OF NORRIS' 'ANCIENT CORNISH DRAMA' (1900) comp. Whitley Stokes. *Archiv für celtische Lexikographie* I 161-174.

CPNE CORNISH PLACE NAME ELEMENTS (1985) comp. Oliver Padel. Nottingham: English Place Name Society.

CPNL CORNISH PLACE NAMES AND LANGUAGE (1995) comp. Craig Weatherhill. Wilmslow: Sigma.

CSCDP CELTIC SURNAMES IN CORNWALL, THEIR DISTRIBUTION AND POPULATION IN 1953, THEIR ORIGINS, HISTORY AND ETYMOLOGY (1970) comp. R. Blewett. Unpublished.

CWOT CORNISH WORDS OCCURRING IN TREGEAR MS. (1950) Comp. Robert Morton Nance. In Robert Morton Nance (1950) "The Tregear Manuscript" *Old Cornwall* IV:2 429-4.

DEL [A] DICTIONARY OF THE ENGLISH LANGUAGE (1755) comp. Samuel Johnson. London: W. Strachan for J. and P. Knapton, T. and L. Longman C. Hitch and L. Hawes, A. Millar

and R. and J. Dodsley.

DL DICTIONARY OF LEXICOGRAPHY (1998) comp. R.R.K. Hartmann & Gregory James. London: Routledge.

DLP2 A DICTIONARY OF LINGUISTICS AND PHONETICS 2nd edn. (1985) comp. David Crystal. Oxford: Basil Blackwell.

EBED ELEMENTARY BRETON - ENGLISH DICTIONARY: GERIADURIG BREZHONEG – SAOZNEG (1979) comp. R. Delaporte. Cork: Cork University Press.

ECCED AN ENGLISH-CORNISH AND CORNISH-ENGLISH DICTIONARY (1978) comp. Robert Morton Nance. Redruth: Cornish Language Board.

ECD1 AN ENGLISH-CORNISH DICTIONARY (1887) comp. F. Jago. London: Simpkin Marshall.

ECD2 AN ENGLISH CORNISH DICTIONARY (1934) comp. Robert Morton Nance & A.S.D. Smith. St. Ives: The Federation of Old Cornwall Societies.

ECD3 AN ENGLISH-CORNISH DICTIONARY (1952) comp. Robert Morton Nance. N.p.: Federation of Old Cornwall Societies.

ECD4 ENGLISH CORNISH DICTIONARY: GERLYVER SAWSNEK-KERNOWEK (2000) comp. Nicholas Williams. Dublin: Everson Gunn Teoranta.

ECWD ENGLISH-CORNISH-WELSH DICTIONARY (n.d.) comp. T. Eurwedd Williams. In two volumes; unpublished MSS.12514 and 12515 in the National Library of Wales.

ELL ENGLISH-LATIN LEXICON (n.d.) comp. Aelfric, Abbot of Eynsham. St. John's College, Oxford, 154 (MS.0).

FNWP THE FIELD-NAMES OF WEST PENWITH (1990) comp. P.A.S. Pool. Hayle: The Author.

GAB2 GLOSSARIUM ANTIQUITATUM BRITANNICARUM: SIVE SYLLABU ETYMOLOGICUS ANTIQUITATUM VETERIS BRITANNIAE ATQUE IBE TEMPORIBUS ROMANORUM / AUCTORE WILLIELMO BAXTER ... ACCEDUNT ... EDUARDI LUIDII ... . DE FLUVIORUM, MONT URBIUM, &C. IN BRITANNIA NOMINIBUS, ADVERSARIA POSTH. 2nd ed. (1733) comp. William Baxter. Londini: T. Woodward.

GAKB GERIADUR ARNEVEZ KERNAOUEK-BREZHONEK (1989) comp. Tim Saunders. Roazhon: Imbourc'h.

GCDBM A GLOSSARY TO THE CORNISH DRAMA 'BEUNANS MERIASEK' (1900) comp. Whitley Stokes. *Archiv für celtische Lexikographie* I 100-142.

GCN GLOSSARY OF CORNISH NAMES (1871) comp. J. Bannister. London: Williams and Norgate.

GCPN A GUIDE TO CORNISH PLACE NAMES (2nd edn.1963a) comp. Robert Morton Nance. n.p.: Federation of Old Cornwall Societies.

GCSW A GLOSSARY OF CORNISH SEA-WORDS (1963) comp. Robert Morton Nance, ed. P.A.S. Pool. . n.p.: Federation of Old Cornwall Societies.

GK GEIRLYFR KYRNWEIG Edward Lhuyd. National Library of Wales, Llanstephan MSS.84.

GKK GERLYVER KERNEWEK KEMMYN: AN GERLYVER MEUR, KERNEWEK-SOWSNEK (1993) Ken George. N.p. Kesva an Tavas Kernewek.

GKKDS GERLYVER KERNEWEK KEMMYN: DYLLANS SERVADOW: SOWSNEK-KERNEWEK (1991) Ken George. N.p.: Kesva an Taves Kernewek.

GN Y GEIRIADUR NEWYDD: THE NEW WELSH DICTIONARY (1952) comp. H. Meurig Evans & W. O. Thomas. Llandybïe: Hughes.

HCS1 HCS2 A HANDBOOK OF CORNISH SURNAMES (1st ed. 1972, 2nd ed.1981) comp. G. Pawley White. Redruth: Dyllansow Truran.

HDHL THE HISTORICAL DICTIONARY OF THE HEBREW LANGUAGE: 1907- (1979) comp. Ze'ev Ben-Hayyim. Tel-Aviv: Tel-Aviv University.

ICPN AN INTRODUCTION TO CORNISH PLACE NAMES (1969) comp. P.A.S. Pool. Penzance: The Author.

LCB LEXICON CORNU-BRITANNICUM - GERLYVR CERNEWEC (1865) comp. R. Williams. London: Trubner.

LK AN LHADYMER AY KERNOU (1750) comp. William Hals. National Library of Wales MSS.1662.

NC NAMES FOR THE CORNISH - 300 CORNISH CHRISTIAN NAMES (1970) comp. C. Bice. Padstow: Lodenek Press.

NCED A NEW CORNISH-ENGLISH DICTIONARY (1938) comp.

Robert Morton Nance. St. Ives: The Federation of Old Cornwall Societies.

NPDMC A NEW PRACTICAL DICTIONARY OF MODERN CORNISH: PART TWO: ENGLISH-CORNISH (1998) R.R.M. Gendall. Menheniot: Teere ha Tavaz.

NSCD THE NEW STANDARD CORNISH DICTIONARY: AN GERLYVER KRES (1998) Ken George. N.p.: Cornish Language Board.

NWEW THE NEW WORLD OF ENGLISH WORDS (1658) comp. Edward Phillips. London: printed for J. Phillips.

OALD3 OXFORD ADVANCED LEARNER'S DICTIONARY OF CURRENT ENGLISH (3rd ed. 1987) edit. A.S. Hornby with A.P. Cowie, A.C. Gimson. Oxford: OUP.

OALD4 OXFORD ADVANCED LEARNER'S DICTIONARY OF CURRENT ENGLISH (4th edn. 1989) edit. A. Cowie. Oxford: OUP.

OTPNC ONE THOUSAND PLACE NAMES OF CORNWALL (n.d.) comp. E. Chirgwin. Privately published.

PCB PATRONYMICA CORNU-BRITANNICA OR THE ETYMOLOGY OF CORNISH SURNAMES (1870) comp. R. Charnock. London: Longman.

PDCPN A POPULAR DICTIONARY OF CORNISH PLACE NAMES (1988) comp. Oliver Padel. Penzance: Alison Hodge.

PDMC A PRACTICAL DICTIONARY OF MODERN CORNISH: PART ONE CORNISH-ENGLISH (1997) R.R.M. Gendall. Menheniot: Teere ha Tavaz.

PGE PONS GROßWÖRTERBUCH ENGLISCH (1981) comp. P. Terrel, V. Calder-Schnorr, W. Morris & R. Breitsprecher Stuttgart: PONS.

PNC THE PLACE NAMES OF CORNWALL (1928) comp. J.Gover. Unpublished MS 1948 in the Royal Institution of Cornwall.

PNWP1 THE PLACE-NAMES OF WEST PENWITH (1st ed.1973) comp. P.A.S. Pool. Penzance: The Author.

PNWP2 THE PLACE-NAMES OF WEST PENWITH (2nd ed.1985) comp. P.A.S. Pool. Penzance: The Author.

RCDFAAF ROBERT & COLLINS DICTIONNAIRE FRANÇAIS-ANGLAIS ANGLAIS-FRANÇAIS (1987) ed. B.Atkins,

A.Duval & R. Milne. Glasgow: Collins.

SDMC A STUDENTS' DICTIONARY OF MODERN CORNISH: PART 1 ENGLISH-CORNISH (3rd. ed. 1991) comp. R.R.M. Gendall. Menheniot: The Cornish Language Council.

TA [A] TABLE ALPHABETICALL comp. Robert Cawdrey. London: I.R. for E. Weaver.

TCPNE 1,000 CORNISH PLACE-NAMES EXPLAINED (1983) comp. Julian Holmes. Redruth: Dyllansow Truran.

VC VOCABULARIUM CORNICUM British Library Cott. Vesp. A xiv; ff. 7a-10a.

VCL VOCABULARY OF THE CORNISH LANGUAGE, COMPILED, WITH ADDITIONS (1861) comp. Charles Rogers. Bodleian, Oxford, MS Corn d1.

VCBL VOCABULARY OF THE CORNU-BRITISH LANGUAGE (1769) William Borlase. In William Borlase *Observations on the Antiquities Historical and Monumental, of the County of Cornwall* 2nd ed. London: W.Bowyer & J.Nichols.

VVB VOCABULAIRE VIEUX-BRETON (1884) comp. J. Loth. Paris: N.p..

Manuscripts Cited

An Lhadymer ay Kernou compiled by William Hals. National Library of Wales: 1662.

Beunans Meriasek Colophon, "Finitur per dominum Rad Ton anno domini 1504". National Library of Wales: MS. Peniarth 105.

Bilbao Manuscripts Collected by Thomas Tonkin. Biblioteca de la Diputación de la Provincia de Vizcaya, Bilbao, Spain. Photocopy in the Royal Institution of Cornwall.

Black Book of Merthen Royal Institution of Cornwall: Black Book of Merthen.

Bodmin Gospels British Library: Add. 9381.

Charter Endorsement British Library: Add. Cart. 19491.

Common-Place Book of William Gwavas (1710) Royal Institution of Cornwall: Cornish Play, The Creation 1698 - Common-Place Book of William Gwavas.

Diary of Richard Symonds (1644) British Library: Add. 17062.

Enys Collection Manuscripts collected by William Scawen. Cornwall Records Office: MS DDEN 1999.

Exeter Consistory Court Depositions (1569-1572) Royal Institution of Cornwall: Henderson MSS., vol. X p.176.

Glasney Cartulary Cornwall Records Office: MS. Dd R(S)59)

Gwavas Manuscripts Collected by William Gwavas. British Library: Add. MSS. 28,554.

Gwreans an Bys Collophon: "William Jordan: the XIIth of August 1611". Bodleian: 219,

Jottings by William Gwavas on reverse of a legal document. Royal Institution of Cornwall.

Keigwin Manuscripts Bodleian: Gough Cornwall 3-4.

Lhuyd's Phonetically Spelled Transcript of James Jenkins' Verses Bodleian: 10712.

Manuscript Belonging to Lhuyd Bodleian: 10,714.

Manuscript of Nicholas Boson's Royal Institution of Cornwall.

Mems. of the Cornish Tongue Manuscripts collected by William Borlase (1749). Cornwall Records Office: DDEN 2000.

Observations on a Manuscript Entitled Passio Christi... by William Scawen. British Library: 628.k.17

Ordinalia Bodleian: 791.

Origo Mundi First play of the Ordinalia. Bodleian: 791, ff. 1a-26a.

Oxoniensis Posterior Bodleian: 572.

Pascon Agan Arluth British Library: Harleian N. 1782.

Passio Domini Second play of the *Ordinalia*. Bodleian: 791, ff. 28a-55b.

Penzance Manuscript Morrab Library, Penzance.

Prophetia Merlini, Joannis Cornubiensis. Vatican: Cod. Ottobonianus Lat. 1474.

Resurrexio Domini Third play of the *Ordinalia*. Bodleian 791, ff. 57a-82a.

Scawen Manuscripts Collected by William Scawen. British Library: Add. MSS 33420.

Smaragdus's Commentary on Donatus Paris Bibliotheque Nat.: MS.Lat. 13029.

Star Chambers Henry VIII, 8/171-175. Cited by Loth (1911b: 443)

Tonkin MSS B Royal Institution of Cornwall: Tonkin MSS. B.

Tonkin MSS H Royal Institution of Cornwall: Tonkin MSS. H.

Tregear Homilies British Library: Add. MS. 46397.

William Bodinar's Letter (1776) Letter to Daines Barrington. Manuscript in the custody of the Society of Antiquities of London.

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.

William Hals' History British Library: Add. 29762.

Software Cited

Linguistica Developed by John Goldsmith, Department of Linguistics, University of Chicago. Available for download from http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000/linguistica2001.exe .

Multiconcord Developed by David Woolls, University of Birmingham. Birmingham: CFL Software Development.

Oxford Concordance Program Oxford: Oxford University Press.

ParaConc Developed by Michael Barlow, Department of Linguistics, Rice University. New York: Athelstan.

Wordsmith Tools Developed by Mike Scott. Oxford: Oxford University Press.

Xcorpus Developed by Patrice Bonhomme & Laurent Romary. Available for download from http:// www.loria.fr/projets/XCorpus/download.html .

Other Works Cited

Aarts, J., P. de Haan & N. Oostdijk eds. (1993) *English Language Corpora: Design, Analysis and Exploitation* Papers from the 13th International Conference on English Language Research on Computerized Corpora, Nijmegen 1992. Amsterdam: Rodopi.

Allin-Collins, R. St. V. (Hal Wyn) (1927) *Cornish Grammar and Supplement to "Some Short Stories in the Cornish Language"* n.p.: the author.

Allin-Collins, R. St. V. (Hal Wyn) (1930) "Is Cornish Actually Dead?" *Zeitschrift für celtische Philologie* XVIII 287-292.

Atkins, B.T.S & Antonio Zampolli eds. (1994) *Computational Approaches to the Lexicon* Oxford: Oxford University Press.

Bakere, Jane A. (1980) *The Cornish Ordinalia: A Critical Study* Cardiff: University of Wales Press.

Baldinger, K. (1971) "Semasiologie und Onomasiologie im zweisprachigung Wörterbuch" *Interlinguistica - Sprachvergleich und Übersetzung. Festschrift zum 60. Geburtstag von Maria Wandruszka* Tübingen, 384-96.

Barlow, Michael (1995) "A Guide to ParaConc" http://www.ruf.rice.edu/~barlow/pc.html .

Barnhart, C.L. (1975) "Problems in Editing Commercial Monolingual Dictionaries" in F.W. Householder & S. Saporta eds. (1975).

Barrington, Daines (1776) "Mr. Barrington on some Additional Information Relative to the Continuance of the Cornish Language: In a Letter to John Lloyd Esq. F.A.S." *Archaeologia* V. London: Society of Antiquities, 81-86.

Batori, S., W. Lenders & W. Putschke eds. (1989) *Computational Linguistics: an International Handbook on Computer Oriented Language Research and Applications* Berlin: Walter de Gruyter.

Bauer, L. (1988) *Introducing Linguistic Morphology* Edinburgh: Edinburgh University Press.

Béjoint, H. (1981) "The Foreign Student's Use of Monolingual English Dictionaries: a Study of Language Needs and Reference Skills" *Applied Linguistics* II:3 207-222.

Béjoint, H. (1994) *Tradition and Innovation in Modern English Dictionaries* Oxford Studies in Lexicography and Lexicology. Oxford: Clarendon Press.

Benson, M. (1989) "The Structure of the Collocational Dictionary" *International Journal of Lexicography* II:1 1-14.

Berber Sardinha, A. P (1996) "Review: WordSmith Tools" *Computers & Texts* XII 19.

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).

Bodinar, William (1776) Letter to Daines Barrington. Manuscript in the custody of the Society of Antiquities of London.

Bogaards, P. (1990) “Où cherche-t-on dans le dictionnaire?” *International Journal of Lexicography* III:2 79-102.

Bonaparte, L. (1861) *Cambrian Journal* 30th Nov. 1861.

Bonaparte, L. (1866) *Some Observations on the Rev. R. Williams’ Preface to his Lexicon Cornu-Britannicam.*

Bonhomme, Patrice & Laurent Romary (1997) *XCORPUS - Version 0.2: A Corpus Toolkit Environment: User Manual* http://www.loria.fr/projets/XCorpus/manual/ .

Bonner, Edmund (1555) *A profitable and necessarye doctryne, with certayne homelies adioyned thervnto set forth by the reuerende father in God, Edmonde byshop of London, for the instruction and enformation of the people beynge within his Diocesse of London, of his cure and charge. B.L.* London: Iohannis Cawodini.

Boorde, A. (n.d.) *Fyrst Boke of the Introduction of Knowledge* London: William Copland.

Boorde, A. (1555) *The Fyrst Boke of the Introduction of Knowledge* London: In fleetestrete, at the signe of the Rose Garland.

Borlase, G. (1733) Copy of some papers of Gwavas in the Royal Institution of Cornwall.

Borlase, W. (1748) “Mems. of Cornish Tongue” Cornwall Record Office DDEN 2000.

Borlase, W. (1749) Letter to William Stukeley in P. Pool (1966: 11).

Borlase, W. (1769) *Observations on the Antiquities Historical and Monumental, of the County of Cornwall* 2nd ed. London: W.Bowyer & J.Nichols.

Botha, W. (1992) “The Lemmatisation of Expressions in Descriptive Dictionaries” in H. Tommola et al. eds. 465-72.

Bottrell, W. (1870) *Traditions and Hearthside Stories of West Cornwall* Vol. I. Penzance: The Author.

Brome, Richard (1632) *The Northern Lasse: A Comoedie* London: A. Mathewes, sold by N. Vavasour.

Brown, Wella (1984) *A Grammar of Modern Cornish* Saltash: The Cornish Language Board.

Brown, Wella (1993) *A Grammar of Modern Cornish* 2nd ed. np. The Cornish Language Board.

Burchfield, R. ed. (1987) *Studies in Lexicography* Oxford: Clarendon Press.

Burgess, A. (1964) *Language Made Plain* London: English Universities Press.

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.

Campanile, H. Enrico (1963) “Un framento scenico Medio-Cornico” *Studi e Saggi Linguistici* III 60-80.

Carew, Richard (1602) *The Survey of Cornwall* London: Printed by S. S[tafford] for Iohn Iaggard.

Catford, J. (1967) “Translation and Language Teaching” *Linguistic Theories and their Application* London: International Association of Publishers of Applied Linguistics 125-146.

Chubb, Ray, Richard Jenkin & Graham Sandercock (2001) *The Cornish Ordinalia: First Play: Origo Mundi* n.p.: Agan Tavas.

Clocksin, W.F. & C.S. Mellish (1981) *Programming in Prolog* Berlin: Springer-Verlag.

Combellack-Harris (1985) *A Critical Edition of Beunans Meriasek* University of Exeter, PhD Thesis.

Combellack-Harris (1988) *The Camborne Play: A Verse Translation of Beunans Meriasek* Redruth: Dyllansow Truran.

Covington, Michael A. (1994) *Natural Language Processing for Prolog Programmers* Englewood Cliffs, New Jersey: Prentice Hall.

Cowie, A.P. (1981) “The Treatment of Collocations and Idioms in Learners’ Dictionaries” *Applied Linguistics* II:3 223-35.

Crawford, T.D. (1980) “The Composition of the Cornish Ordinalia” *Old Cornwall* IX:3 145-153.

Crystal, David (1985) *A Dictionary of Linguistics and Phonetics* 2nd edn.. Oxford: Basil Blackwell.

Cuillandre, J. (1931) “Contributions à l’étude des textes corniques” *Revue celtique* XXXXVIII 1-41.

Cuillandre, J. (1932) “Contributions à l’étude des textes corniques” *Revue celtique* XXXXIX 109-131.

Curley, Michael J. (1982) “A New Edition of John of Cornwall’s ‘Prophetia Merlini’” *Speculum* LVII:2 217-249.

Davies, W. (1939) *Cornish MSS in the National Library of Wales* Pamphlet in the National Library of Wales.

Edwards, Ray ed. (1993) *Pascon Agan Arluth: Passhyon agan Arloedh* Sutton Coldfield: Kernewek dre Lyther.

EKOS Ltd. & SGRÙD Research (2000) “An Independent Academic Study on Cornish” Government Office for the South West http://www.gosw.gov.uk/publications/Cornish/studycover.htm .

Eyes, E. & G. Leech (1992) “Progress in UCREL Research: Improving Corpus Annotation Practices” in Aarts, de Haan & Oostdijk (1993) 123-43.

Firth, J.R. (1957) “Applications of General Linguistics” *Transactions of the Philological Society* 1-14.

Fleuriot, L. (1974) “Les fragments du texte brittonique de la Prophetia Merlini” *Études celtiques* XIV 31-56.

Flowerdew, Lynne & Anthony K.K. Tong eds. (1994) *Entering Text* Hong Kong: The Language Centre, The Hong Kong University of Science and Technology.

Fowler, D. (1961) “The Date of the Cornish Ordinalia” *Medieval Studies* XXIII 91-125.

Francis, W. (1980) “A Tagged Corpus - Problems and Prospects” *Studies in English Linguistics* eds. S. Greenbaum, G. Leech & J. Svartvik. London: Longman 192-209.

Fudge, Crysten (1982) *The Life of Cornish* Redruth: Dyllansow Truran.

Gal, Annie, Guy Lapalme, Patrick Saint-Dizier & Harold Somers (1991) *Prolog for Natural Language Processing* Chichester: John Wiley & Sons.

Garside, Roger, Geoffrey Leech & Anthony McEnery eds. (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora* Harlow: Addison Wesley Longman.

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

Geeraerts, D. (1989) “Principles of Monolingual Lexicography” in Hausmann et al. eds. 287-96.

Gellerstam, M., J. Järborg, S. G. Malmgren, K. Norén, L. Rogström & C.R. Papmehl eds. (1996) *Euralex 96: Proceedings of the 7th Euralex International Congress on Lexicography - Euralex 96 Proceedings I-II Submitted to the International Congress on Lexicography in Göteborgm, Sweden* Gothenburg: Gothenburg University, Department of Swedish.

Gendall, R.R.M. (1991) *The Pronunciation of Cornish* 2nd ed. Menheniot: Teere ha Tavaz.

George, Ken (1983) “A Computer Model of Sound Changes in Cornish” *Journal of the Association of Literary and Linguistic Computing* IV 39-48.

George, Ken (1984) *A Phonological History of Cornish* Univ. Western Britanny (Brest): Thesis.

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

George, Ken (1991) “The Nouns Suffixes -ter/-der, -(y)ans and -neth in Cornish” *Études celtiques* XXVIII 203-212.

Gilbert, Davies ed. (1826) *Mount Calvary; or the History of the Passion, Death, and Resurrection of our Lord and Saviour Jesus Christ* London: Nichols and Son.

Gilbert, Davies ed. (1827) *The Creation of the World with Noah’s Flood*

London: J.B. Nichols.

Gilbert, Davies (1838) *The Parochial History of Cornwall: Ffounded on the Manuscript Histories of Mr. Hals and Mr. Tonkin, with Additions and Various Appendices* London: J.B. Nichols.

Gorcy, G. (1989) "Différenciation des significations dans le dictionnaire monolingue: problèmes et méthodes" in Hausmann et al. eds. (1991) 905-17.

Graves, E. ed. (1962) *The Old Cornish Vocabulary* Ann Arbor: University Microfilms Inc..

Guo Jin (1996) "An Efficient and Complete Algorithm for Unambiguous Word Boundary Identification" http//:sunzi.iss.nus.sg:1996/guojin/papers/acbci/acbci.rtf .

Guo, Cheng-ming ed. (1995) *Machine Tractable Dictionaries: Design and Construction* Norwood, N.J.: Ablex.

Halliday, F.E. ed. (1953) *Richard Carew of Anthony: The Survey of Cornwall etc.* London: Melrose.

Halliday, M.A.K. (1956) "Grammatical Categories in Modern Chinese" *Transactions of the Philological Society* 177-224.

Halliday, M.A.K. (1961) "Categories of the Theory of Grammar" *Word* XVII 241-92.

Halliday, M.A.K. (1994) *An Introduction to Functional Grammar* 2nd ed. London: Edward Arnold.

Hamp, Eric P. (1958-1959) "Middle Welsh, Cornish and Breton Personal Pronominal Forms" *Études celtiques* VIII 394-401.

Harris, Markham (1969) *The Cornish Ordinalia: A Medieval Dramatic Trilogy* Washington: University of America Press.

Harris, Phyllis Pier ed. and trans. (1964) *Origo Mundi: First Play of the Cornish Mystery Cycle, The Ordinalia: A New Edition* University of Washington: PhD Thesis; Ann Arbor: University Microfilms, (1964).

Hartmann, R.R.K. ed. (1983) *Lexicography: Principles and Practice* Applied Language Studies. London: Academic Press.

Hartmann, R.R.K. ed. (1984) *LEXeter '83 Proceedings* Papers from the International Conference on Lexicography at Exeter, 9-12 September 1983. Tübingen: Niemeyer.

Hartmann, R.R.K. & Gregory James (1998) *Dictionary of Lexicography* London: Routledge.

Hausmann, F. J., O. Reichmann, H.E. Wiegand & L. Zgusta eds. (1989-1991) *Wörterbücher/Dictionaries/Dictionnaires: An International Encyclopedia of Lexicography* Handbücher zur Sprachund Kommunikationswissenschaft 5.1, 5.2, 5.3. Berlin: Walter de Gruyter.

Hausmann, F.J. & H.E. Wiegand (1989-1991) “Component Parts and Structures of General Monolingual Dictionaries: A Survey” in F.J. Hausmann et al. eds. (1989-1991).

Hausmann, F.J. (1989-1991) “Le Dictionnaire de collocations” in Hausmann et al. eds. (1989-1991) 1010-19.

Hawke, Andrew (1979) “A Lost Manuscript of the Cornish Ordinalia” *Cornish Studies* Series 1. VII, 45-60.

Hawke, Andrew (1981) “The Manuscripts of the Cornish Passion Poem” *Cornish Studies* Series 1. IX 23-28.

Hawke, Andrew (1982) *The Cornish Dictionary Project* Redruth: Institute of Cornish Studies.

Hawke, Andrew (2001) “A Rediscovered Cornish-English Vocabulary” *Cornish Studies* Series 2. IX 83-104.

Haywood, N. (1982) *Studies in the Historical Grammar of Cornish* Univ. of Oxford: dissertation.

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

Hickey, R. (1994) “Applications of Software in the Compilation of Corpora” M. Kytö, M. Rissanen & S. Wright eds. (1994).

Hoblyn, W. Treffry (1936) “ ‘In English & not in Cornowok’: From the Consistory Court Depositions, 1569-1572 (Special Commission)” *Old Cornwall* II:11 p.11.

Hobson Matthews, J. (1892) *A History of the Parishes of Saint Ives, Lelant, Towednack and Zennor, in the County of Cornwall* London: Elliot Stock.

Hooper, E.G.R. ed. (1972a) *Passyon agan Arluth: Cornish Poem of the Passion in Unified Cornish by R. Morton Nance and A.S.D. Smith* n.p.: Kesva an Tavas Kernewek.

Hooper, E.G.R. (1972b) Letter to Berresford Ellis. In P. Berresford Ellis (1974) 192.

Hooper, E.G.R. ed. (1985) *Gwryans an Bys or The Creation of the World: As Written by William Jordan, A.D. 1611: In Unified Spelling with Amended Translation by R. Morton Nance (Mordon) and A.S.D. Smith (Caradar)* Redruth: Dyllansow Truran.

Householder, F.W. (1975) “Summary Report” in F.W. Householder & S. Saporta eds. (1975).

Householder, F.W. & S. Saporta eds. (1975) *Problems in Lexicography* Bloomington, Ind.: Indiana University Press.

Hudson, R.A. (1988) “The Linguistic Foundations for Lexical Research and Dictionary Design” *International Journal of Lexicography* I:4 287-312.

Hull, Vernam (1958-1959) “Celtic Manuscripts in Spain and in Portugal” *Zeitschrift für celtische Philologie* XXVII 227-229.

Ilson, R.F. (1984) “The Communicative Significance of some Lexicographic Conventions” in Hartmann ed. (1984) 80-6.

Ilson, R.F. (1988) “Contributions to the Terminology of Lexicography” in *ZüriLEX ‘86 Proceedings* ed. M. Snell-Hornby. Tubingen: EURALEX.

Jackson, Kenneth Hurlstone (1967) *A Historical Phonology of Breton* Dublin: Dublin Inst. for Advanced Studies.

Jenner, Henry (1877) “An Early Cornish Fragment” *The Athenaeum* 2614 698-699.

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

Jenner, Henry (1912) “Description of Cornish Manuscripts 1: The Borlase MSS” *Journal of the Royal Institute of Cornwall XIX* 162-76.

Jenner, Henry (1915-1916) “Description of Cornish Manuscripts II: The fourteenth-century Charter Endorsement, Brit. Mus. Add. Ch. 19491” *Journal of the Royal Institution of Cornwall* XX 46-47.

Jenner, Henry (1925) “The Cornish MSS in the Provincial Library at Bilbao, Spain” *Journal of the Royal Institution of Cornwall* XXI 421-37.

Jenner, Henry (1928) “King Teudar” *Tre, Pol and Pen* London: The Cornish Association.

Jenner, Henry (1929) “Some Miscellaneous Scraps of Cornish” *96th Annual Report of the Royal Cornwall Philological Society* VI:3 238-55.

Jones, R. & S. Sondrup (1989) “Computer Aided Lexicography: Indexes and Concordances” in Batori, Lenders & Putschke eds. 490-509.

Jordan, William (1611) *Gwreans an Bys* Unpublished manuscript Bodleian 219.

Kennedy, Graeme (1998) *An Introduction to Corpus Linguistics* London: Longman.

Kennedy, Neil (2001) “Review Article: Gerlyver Sawsnek-Kernowek” *Cornish Studies* Series 2. IX.

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.

Knowles, F. (1983) “Towards the Machine Dictionary: Mechanical Dictionaries” in Hartmann ed. (1983) 181-93.

Kromann, H., T. Riiber & P. Rosbach (1991) “Theory of Bilingual and Multilingual Lexicography I: Principles and Components” in F.Hausmann et al. eds. (1991).

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.

Landau, Sydney (1989) *Dictionaries: The Art and Craft of Lexicography* Cambridge: CUP.

Leech, Geoffrey (1981) *Semantics: the Study of Meaning* 2nd ed. London: Penguin.

Leech, Geoffrey (1997) "Grammatical Tagging" Garside, Leech & McEnery eds. (1997).

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

Lewis, Henry (1990) *Handbuch des Mittelkornischen / Henry Lewis; deutsch Bearbeitung von Stefan Zimmer; mit einem bibliographischen Anhang von Andrew Hawke* trans. Stefan Zimmer. Innsbrucker Beitrage zur Sprachwissenschaft Innsbruck: Institut fur Sprachwissenschaft. Translation of: Henry Lewis (1923) *Llawlyfr Cernyweg Canol* Wrecsam: n.p..

Lindsay, W.M. (1897) "A Welsh (Cornish?) gloss in a Leyden MS." *Zeitschrift für celtische Philologie* I , 361.

Long, C. ed. (1856) *Diary of the Marches of the Royal Armay during the Great Civil War. Kept by Richard Symonds (1644)* London: Camden Society.

Lorentzen, H. (1996) "Lemmatisation of Multi-word Lexical Units: In Which Entry?" in M. Gellerstam et al. eds. 415-21.

Loth, J. (1893a) "Les gloses de l'Oxoniensis posterior sont-elles corniques" *Revue celtique* XIV, 70.

Loth, J. (1893b) "Les mots 'druic', 'nader', dans le Vocabulaire cornique" *Revue celtique* XIV, 301-304.

Loth, J. (1897) "Etudes corniques" *Revue celtique* XVII, 401-422.

Loth, J. (1900) "Cornique moderne" *Archiv für celtische Lexikographie* I, 224-229.

Loth, J. (1902a) "Etudes corniques II: Textes inédits en cornique moderne" *Revue celtique* XXIII, 173-200.

Loth, J. (1902b) "Etudes corniques III/IV: Remarques et corrections au Lexicon Cornu-Britannicum de Williams" *Revue celtique* XXIII, 236-302.

Loth, J. (1903) "Études corniques V: Les dix commandements de Dieu" *Revue celtique* XXIV, 1-10.

Loth, J. (1905) "Etudes Corniques VI: Corrections a divers textes corniques" *Revue celtique* XXVI, 218-267.

Loth, J. (1906) “Le cornique moderne: à propos d’un livre de M. Henry Jenner” *Revue celtique* XXVII, 93-101.

Loth, J. (1907a) “Les Gloses à Smaragdus” *Archiv für celtische Lexikographie* III, 249-256.

Loth, J. (1907b) “Etymologies diverses” *Archiv für celtische Lexikographie* III, 257-265.

Loth, J. (1907c) “Les Gloses à Smaragdus” *Revue celtique* XXV, 215-6.

Loth, J. (1911a) “Cornoviana: Les bretons en cornwall au commencement di XVIe siècle” *Revue celtique* XXXII, 290-295.

Loth, J. (1911b) “Cornoviana: Les bretons en cornwall, note additionelle: Une phrase inèdite en moyen cornique et un mot rare: Un usage des iles scilly” *Revue celtique* XXXII, 442-445.

Loth, J. (1913) “Cornoviana” *Revue celtique* XXXIV, 176-181.

Loth, J. (1914) “Questions de grammaire: ... Le cornique moderne: traits principaux de sa phonetique et de sa syntaxe” *Revue celtique* XXXV, 143-164.

Loth, J. (1917-1919) “Questions de grammaire: ... Le cornique moderne; l’orthographie et les sons” *Revue celtique* XXXVII, 151-211.

Madan, F. & H.H.E. Craster (1922) *A Summary Catalogue of Western Manuscripts in the Bodleian Library at Oxford* Vol. II:1 Oxford.

Makkai, A. (1980) “Theoretical and Practical Aspects of an Associative Lexicon for 20th Century English” L. Zgusta (1980) ed..

Malkiel, Y. (1975) “A Typological Classification of Dictionaries on the Basis of Distinctive Features” F.W. Householder & S. Saporta eds. (1975).

Marinone, N. (1981) “A Project for a Latin Lexical Data Base” in Zampolli and Cappelli eds. (1983) 175-8.

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

Martin, S.E. (1967) “Selection and Presentation of Ready Equivalents in a Translation Dictionary” *Problems in Lexicography.* 2nd ed. edit. F.W. Householder & Sol Saporta. Bloomington, Ind. Indiana University.

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).

Mathiot, M. (1967) “The Place of the Dictionary in Linguistic Description: Problems and Implications” *Language* XXXXIII, 703 *ff.*.

Matoré, G. (1968) *Histoire des dictionnaires français* Paris: Larousse.

McArthur, T. (1986) *Worlds of Reference: Lexicography, Learning and Language from the Clay Tablet to the Computer* Cambridge: Cambridge University Press.

M^{c}Carthy, M. ed. (1988) "Naturalness in Language" *ELR Journal* II.

McEnery, Tony & Andrew Wilson (1996) *Corpus Linguistics* Edinburgh: Edinburgh University Press.

Mencken, H.R. (1923) *The American Language* New York: Alfred A. Knopf (3rd ed.).

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

Milles, J. (1753) "Copy of the Cornish Vocabulary in the Cotton: Library London Copy'd by the Revd Dr. Jer: Milles Chantor of the Church of Exeter, 1753" in W. Borlase (1748).

Mills, Jon (1992) *Componential Analysis of Meaning from Computer Concordances of Middle to Late Cornish* University of Exeter: Dissertation.

Mills, Jon (1999) "Reconstructive Phonology and Contrastive Lexicology: Problems with the 'Gerlyver Kernewek Kemmyn'" *Cornish Studies* Series 2.VII ed. Philip Payton 193-218.

Morris Jones, Bob (1983-1984) "Automatic Tagging and Lemmatization: Using SPITBOL as an Aid in Lexical and Grammatical Analysis of Welsh Texts" *Studia Celtica* XVIII/XIX 287-310.

Morton Nance, Robert (n.d.) Unpublished MSS in the Royal Institution of Cornwall.

Morton Nance, Robert (1922-1925) "John Davey of Boswednack and his Cornish Rhyme" *Journal of the Royal Institution of Cornwall* XXII, 146-153.

Morton Nance, Robert (1923) "Celtic Words in Cornish Dialect" Report Royal Cornwall Polytechnic Society IV Falmouth: Royal Cornwall Polytechnic Society.

Morton Nance, Robert (1928) "Andrew Boorde on Cornwall" *Journal of the Royal Institution of Cornwall* XXII:3, 366-381.

Morton Nance, Robert (1929) *Cornish for All* n.p.: Federation of Old Cornwall Societies.

Morton Nance, Robert (1932) "The Charter Endorsement in Cornish" *Old Cornwall* II:4, 34-36.

Morton Nance, Robert (1934-1936) "*Pascon Agan Arluth*" *Kernow* 1-14.

Morton Nance, Robert (1945) "Celtic Personal Names of Cornwall" *Old Cornwall* IV, 10-17 and 61-8.

Morton Nance, Robert (1947) "New Light on Cornish" *Old Cornwall* IV:6,

214-216.

Morton Nance, Robert (1949) "A Cornish Poem Restored" *Old Cornwall* IV:10, 368-371.

Morton Nance, Robert (1950) "The Tregear Manuscript" *Old Cornwall* IV:2, 429-4.

Morton Nance, Robert (1951) "More about the Tregear Manuscript" *Old Cornwall* V:1, 21-27.

Morton Nance, Robert (1954) "Cornish Words in the Tregear MS" *Zeitschrift für celtische Philologie* XXIV, 1-5.

Morton Nance, Robert & A.S.D. Smith (n.d.) "Ordinale de Vita Sancti Mereadoci, Episcopi et Confessoris: Bewnans Meryasek" in the Royal Institution of Cornwall.

Morton Nance, Robert & A.S.D. Smith (1959) *Gwryans an Bys* Padstow: Federation of Old Cornwall Societies.

Morton Nance, Robert & A.S.D. Smith (1963-1968) "Passyon agan Arluth" *An Lef Kernewek* 81-103.

Morton Nance, Robert & A.S.D. Smith (1966) *St. Meriasek in Cornwall* Extracts from the Cornish Texts in unified spelling with amended translation. No. 1. N.p.: Federation of Old Cornwall Societies.

Morton Nance, Robert & A.S.D. Smith (1969) *An Venen ha'y Map* Extracts from the Cornish Texts in unified spelling with amended translation. No. 7. N.p.: The Cornish Language Board.

Morton Nance, Robert & A.S.D. Smith (1974) *Sylvester ha'n Dhragon* Extracts from the Cornish Texts in unified spelling with amended translation. No. 3. N.p.: The Cornish Language Board.

Mugdan, J. (1991) "Information on Inflectional Morphology in the General Monolingual Dictionary" in F.Hausmann et al. eds. (1991).

Muller, C. (1977) *Principes et méthodes de statistique lexicale* Paris: Hachette.

Murdoch, Brian (1979) *Institute of Cornish Studies, Special Bibliography No. 5: The Medieval Cornish Poem of the Passion* Redruth: Institute of Cornish Studies.

Murdoch, Brian (1993) *Cornish Literature* Cambridge: D.S. Brewer.

Neuss, Paula (1971) *The Creacion of the World: A Critical Edition and Translation* New York: Garland.

Newell, Leonard E. (1995) *Handbook on Lexicography for Philippine and other Languages* Manila: Linguistic Society of the Philippines.

Nida, E. (1958) "Analysis of Meaning and Dictionary Making" in *International Journal of American Linguistics* XXIV, 279-92.

Nida, E. (1976) *The Descriptive Analysis of Words* 2nd edn. Michigan: University of Michigan Press.

Norris, Edwin (1859a) ed. *The Ancient Cornish Drama* Oxford: OUP.

Norris, Edwin (1859b) *Sketch of Cornish Grammar* Oxford: OUP.

O'Keefe, Richard A. (1990) *The Craft of Prolog* Cambridge, Massachusetts: The MIT Press.

Oldwanton, Olyver (1555) *A lyttle treatyse called the Image of Idlenesse conteynynge certeyne matters moued betwene Walter Wedlocke and Bawdin Bachelor. Translated out of the Troyane or Cornyshe tounge by Olyuer Oldwanton and dedicated to the Lady Lust.* London: William Seres dwellynge in Powles Churche yard at the signe of the Hedge hogge.

Olson, Lynette (1997) "Tyranny in 'Beunans Meriasek'" *Cornish Studies* Series 2. V, 52-59.

Ooi, Vincent B. Y. (1998) *Computer Corpus Lexicography* Edinburgh: Edinburgh University Press.

Osselton, N.E. (1995) *Chosen Words: Past and Present Problems for Dictionary Makers* Exeter Linguistic Studies. Exeter: University of Exeter Press.

Palmer, F. (1981) *Semantics* 2nd ed. Cambridge: CUP.

Pan Zaiping & H.E. Wiegand (1987) "Konzeption für das Große Deutsch-Chinesische Wörterbuch" (Zweiter Entwurf) in *Lexicographica* 3.

Partridge, Eric (1963) *The Gentle Art of Lexicography* London: Andre Deutsch.

Payton, Philip (1993) "'A Concealed Envy against the English': A Note on the Aftermath of the 1497 Rebellions in Cornwall" *Cornish Studies* Series 2. I, 4-13.

Pedlar, E. Hoblyn (1859) "Notes on the Names of Places, &c. Mentioned in the Preceding Dramas" in Norris (1859a: 473-514).

Peirce, C.S. (1931-1958) *Collected Papers* vols. 1-8 edited by C. Hartshorne & P. Weiss. Cambridge, Mass.: Harvard University Press.

Penglase, C. (1994) "Authenticity in the Revival of Cornish" *Cornish Studies* Series 2. II. 96-107.

Pennaod, Goulven (1981) *Passyon agan Arluth: Pasion hon Aotrou, barzhoneg kerevek eus ar 15. kantved* Quimper: Preder.

Polwhele, Richard (1816) *The History of Cornwall; Civil, Military, Religious, Architectural, Agricultural, Commercial, Biographical and Miscellaneous New edition .... enlarged.* 7 vols. London: Law & Whittaker.

Pool, P.A.S. & O.J. Padel (1975-1976) "William Bodinar's Letter. 1776" *Journal of the Royal Institution of Cornwall* New Series VII:3, 231-236.

Pool, P.A.S. (1966) "The Borlase - Stukeley Correspondence" *Cornish Archaeology* V, 11.

Pool, P.A.S. (1982) *The Death of Cornish* n.p. Cornish Language Board.

Quentel, P. (1982) "Notes corniques" *Zeitschrift für Celtische Philologie* XXXIX, 195-203.

Radford, Andrew (1988) *Transformational Grammar: A First Course* Cambridge: CUP.

Rawe, Donald (1978) *The Creation of the World* Padstow: Lodenek Press.

Reaney, P. (1960) *The Origin of English Place Names* London: K. Paul.

Rey-Debove, J. (1971) *Etude linguistique et sémiotique des dictionnaires français contemporains* The Hague: Mouton.

Rey-Debove, J. (1989) "Les Systèms de renvois dans la dictionnaire monolingue" in Hausmann et al. eds. 931-7.

Rey, A. (1970) *Littré, l'humaniste et les mots* Paris: Gallimard.

Rey, A. (1972) "Usages, jugements et prescriptions linguistiques" *Langue française* XVI, 4-28.

Rey, A. (1977) *Le Lexique: images at modèles. Du dictionnaire à la lexicologie* Paris: Armand-Collin.

Rey, A. (1982) *Encyclopédies et dictionnaires* Paris: Presses Universitaires de France.

Rissanen J. (1987) *Minimum Description Length Principle* Encyclopaedia of Statistical Sciences Vol. 5. New York: Wiley.

Rissanen, M. (1994) "The Helsinki Corpus of English Texts" *Corpora Across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catherine's College Cambridge, 25-27 March 1993.* Ed. M. Kytö, M. Rissanen and S. Wright. Language and Computers: Studies in Practical Linguistics. Amsterdam: Rodopi.

Robins, R.H. (1987) "Polysemy and the Lexicographer" in R. Burchfield ed. (1987) 52-75.

Sampson, G. (1993) *English for the Computer* Oxford: OUP.

Sandercock, Graham ed. (1982) *The Cornish Ordinalia, Second Play: Christ's Passion: in Modern Cornish with English Translation by R. Morton Nance and A.S.D. Smith* n.p.: Kesva an Tavas Kernewek.

Sandercock, Graham ed. (1984) *The Cornish Ordinalia, Third Play: Resurrection: in Modern Cornish with English Translation by R. Morton Nance and A.S.D. Smith* n.p.: Kesva an Tavas Kernewek.

Sandercock, Graham ed. (1989) *The Cornish Ordinalia, First Play: Origo Mundi Lines 1-465: in modern Cornish with English Translation by R. Morton Nance and A.S.D. Smith* n.p.: Kesva an Tavas Kernewek.

Sandys (1846) *Specimens of Cornish provincial Dialect, collected and arranged by Uncle Jan Trenoodle* London: J. Smith.

Scawen, William (1777) *Observations on and Ancient manuscript entitled Passio Christi, written in the Cornish Language, and now preserved in the Bodleian Library; with an account of the Language, Manners and Customs of the People of Cornwall* London: n.p..

Schnorr, V. (1991) "Problems of Lemmatisation in the Bilingual Dictionary" in F.J. Hausmann et al. eds. (1989-1991).

Schuchardt, H. (1866-68) *Der Vokalismus des Vulgärlateins.* 3 vols. Leipzig: Teubner.

Sebeok, T. (1962) "Materials for a Typology of Dictionaries" *Lingua XI* 363-74.

Sheard, J.H. (1954) *The Words We Use* London: Deutsch.

Sinclair, John (1991) *Corpus, Concordance, Collocation* Oxford: OUP.

Sledd, J.H. & G.J. Kolb (1955) *Dr. Johnson's Dictionary. Essays in the Biography of a Book* Chicago: University of Chicago Press.

Smith, A.S.D. (Caradar) (1972) *Cornish Simplified: Short Lessons for Self-tuition* ed. E.G.R. Hooper. 2nd ed. Redruth: Dyllansow Truran.

Smith, A.S.D. (Caradar) (1984) *Cornish Simplified: Part Two* ed. E.G.R. Hooper (Talek) Redruth: Dyllansow Truran.

Sondrup, S. & C. Inglis (1982) *Konkordanz zu den Gedichten Hugo von Hofmannsthals* Amsterdam: Provo.

Stokes, Whitley ed. (1861) *The Passion: A Middle Cornish Poem* London: Philological Society.

Stokes, Whitley ed. (1863) *Gwreans an Bys: The Creation of the World, a Cornish Mystery, Edited, with a Translation and Notes* Berlin: Asher.

Stokes, Whitley (1870-1872) "The Manumissions of the Bodmin Gospels" *Revue celtique* I, 332-345.

Stokes, Whitley ed. (1872) *Beunans Meriasek: The Life of St Meriasek, Bishop and Confessor: A Cornish Drama* London: Trübner.

Stokes, Whitley (1876-1878) "Cornica" *Revue celtique* III, 85-86.

Stokes, Whitley (1879-1880) "Cornica: The Fragments of a Drama: Cornish Phrases: Poli, Poly" *Revue celtique* IV, 258-264.

Stokes, Whitley ed. (1879) *Old Breton Glosses* Calcutta: The Author.

St. John, Elke & Marc Chattle (1998) "Review of Multiconcord: The Lingua Multilingual Parallel Concordancer for Windows" *ReCALL Newsletter No 13, March 98* .

Svensén, Bo (1993) *Handbok i Lexikografi. Practical Lexicography: Principles and Methods of Dictionary-making* translated from the Swedish by John Sykes and Kerstin Schofield. Oxford: Oxford University Press.

Swanson, D.C. (1975) "The Selection of Entries for a Bilingual Dictionary" in F.W. Householder & S. Saporta eds. (1975).

Symonds, R. (1644) “Diary of the Marches of the Royal Army during the Great Civil War” in C. Long (1856).

Thomas, Charles (1972) “Letter to P. Berresford Ellis” in Berresford Ellis (1974: 194).

Thomas, Jenny & Mick Short eds. (1996) *Using Corpora for Language Research* London: Longman.

Tommola, H., K. Varantola, T. Salmi-Tolonen eds. (1992) *EURALEX ‘92: Proceedings I-II: Papers Submitted to the 5th EURALEX International Congress on Lexicography in Tampere, Finland* Tampere: Tampereen yliopisto.

Tonkin, Thomas (1736) “Letter to William Gwavas” in Bonaparte (1861).

Tonkin, Thomas (1738) “Letter to William Gwavas” in Bonaparte (1861).

Tonkin, Thomas ed. (1811) *Carew’s Survey of Cornwall* London: J. Faulder.

Toorians, Lauran (1991) *The Middle Cornish Charter Endorsement: the making marriage in medieval Cornwall / Middle Cornish text, introduction, translation, commentary and glossary critically edited by Lauran Toorians; with a paleol description of the manuscript by J.P.M. Jansen* Innsbrucker Beitrage zur Sprachwissenschaft. Innsbruck: Institut fur Sprachwissenschaft.

Trench, R.C. (1857) “On some Deficiencies in our English Dictionaries” *Transactions of the Philological Society 1857* 1-70.

Vendryes, J. (1938) “Review of R. Morton Nance ‘A New Cornish-English Dictionary’” *Études celtiques* III, 392-394.

Walker, Donald E., Antonio Zampolli, and Nicoletta Calzolari eds. (1994) *Automating the Lexicon: Research and Practice in a Multilingual Environment* Oxford: Oxford University Press.

Weiner, E. (1994) “The Lexicographical Workstation and the Scholarly Dictionary” in B.T.S. Atkins & A. Zampolli (1994).

Whitaker, John (1804) *The Ancient Cathedral of Cornwall Historically Surveyed* London: John Stockdale.

Wiegand, H. (1984) “Prinzipien und Methoden historischer Lexicographie” in *Sprachgeschichte. Ein Handbuch zur Geschichte der Deutschen Sprache und ihre Erforschung* Handbücher zur Sprach- und Kommunikationswissenschaft ; 2 herausgegeben von W. Besch, O.Reichmann and S.Sonderegger. Berlin: Walter de Gruyter.

Williams, G.P. (1910) “The Preverbal Particle ‘Re’ in Cornish” *Zeitschrift für celtische Philologie* VII, 313-353.

Williams, Nicholas 1997 *Clappya Kernowek: An Introduction to Unified Cornish Revised* n.p.: Agan Tavas.

Williams, N.J.A. (1995) *Cornish Today: An Examination of the Revived Language* Sutton Coldfield: Kernewek dre Lyther.

Williams, N.J.A. (1996) "'Linguistically Sound Principles': The Case against Kernewek Kemmyn" *Cornish Studies* Series 2. III. 64-87.

Williams, N.J.A. (2001) "'A Modern and Scholarly Cornish-English Dictionary': Ken George's 'Gerlyver Kernewek Kemmyn' (1993)" *Cornish Studies* Series 2. IX, 247-311.

Williams, R. (1869) "Cornish Literature" *Archaeologia Cambrensis* XV Third Series, 408-409.

Wrenn, C.L. (1949) *The English Language* London: Methuen.

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

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.

Zeuss, J. (1853) *Grammatica Celtica* Lipsiae: apud Weidmannos.

Zgusta, L. (1971) *Manual of Lexicography* The Hague: Mouton.

Zgusta, L. ed. (1980) *Theory and Method in Lexicography: Western and Non-Western Perspectives* Columbia, North Carolina: Hornbeam Press

Zgusta, L. (1989) "The Influence of Scripts and Morphological Language Types on the Structure of Dictionaries" in F.J. Hausmann et al. eds. 296-305.

Index