Welsh Lexicography and Terminology: past, present and future (I)

Geraint Jones

May I first of all thank you most warmly for your kind invitation to address this conference. My colleague and I are most anxious to exchange experiences and ideas with you in the field of Lexicography believing that there is much that we can learn from the Basque experience.

On such occasions, the first question that I usually have to answer is *where is Wales?* It lies, as some of you know, in the West between England and Ireland and has a population of approximately 3 million people. Roughly 20% of the population speak Welsh as their first language although the percentage in some areas can be as high as 80%. The language is taught as part of the National Curriculum - either as a first or second language - in all the schools of Wales. It is also employed as a medium in an increasing number of primary and secondary schools - the latest figure for the secondary sector being 54.

And what is CANOLFAN BEDWYR? It is a Centre created by University of Wales, Bangor to do three things. Firstly to provide linguistic support for those whose command of the Welsh language is insufficient to be able to take full advantage of University Education. Secondly to conduct research into bilingualism, being especially concerned with different models of bilingual classroom delivery, and lastly to foster Language Engineering which brings me to the topic of the present address.

Whenever I think of *lexicography* I'm reminded of Burgos cathedral. It is a magnificent building, but one that has been added to and adapted over many centuries. Each alteration has drawn inspiration from what was previously accomplished. Lexicographers are equally dependent on their antecedents ever building on the work of previous scholars.

Even the founding father of Welsh lexicography, William Salesbury, whose *Dictionary in Englyshe and Welshe* was published in 1547, drew heavily on the language of the poets and storytellers of previous centuries and on the manuscript glosses made by scribes and scholars. His aim was to ensure that the Welsh language could face the main challenges of the time - the newly established printing press, the challenges of the Renaissance and the English tongue. Welsh could best be developed by providing sound linguistic tools and good, elegant models of language use. Only then could it take its place among the languages of Europe. Such was the thinking behind his dictionary and numerous translations.

The next period of development ran from 1650 - 1850 and embracing a number of diverse contributors like John Walters,Thomas Edwards and William Owen Pugh. Faced with the challenges of a rapidly expanding universe they saw that the language was unable to cope with their accumulating needs. Confronted with such concerns, they were much more prepared than their predecessors *to coin new words.* Some of their surviving creations give us a glimpse of the developments of the age:

'adnoddau/ resources', 'buddsoddi/invest', 'pwyllgor/ committee', 'ffaith/ fact', 'safon/ standard', 'daeareg/ geology', 'amaethyddiaeth/agriculture','nwy/ gas', etc.

It should also be said that many of their coinages failed to gain public favour and consequently met a sad and early death. Lexicographers are ever at the mercy of language users.

During the next period, extending from 1848 to 1903, lexicographers used a more scientific and methodical approach. This is typified by D. Silvan Evans's ambition to compile a *Welsh/English dictionary* - that would favourably compare with the *Oxford English Dictionary.* In contrast to the *team* working at Oxford, Evans worked tirelessly, on his own, for around forty years - publishing his work by instalments from 1887 to 1896. By his death he was only three quarters of the way through the letter 'E'.

The early twentieth century saw such scholars as Sir John Morris Jones, a pragmatic mathematician, rejecting what he considered to be the romantic 19th century approach and returning the language - as they saw it - to its mainstream roots. This happily coincided with the establishment in 1919, by the national University of Wales, of the Board of Celtic Studies. Here the dream of creating the Welsh equivalent of the Oxford English Dictionary was again resurrected. The project was started in 1920 under the direction of the Rev. J Bodfan Anwyl and housed at the National Library of Wales at Aberystwyth - which had earlier been given its charter in 1907. It was a logical choice as it contained the greatest repository of Welsh manuscripts and printed books. The project's aim was to trace the etymology of the Welsh language's vocabulary, in alphabetical order.

The project obviously entailed substantial research, but the number of appointed staff was extremely small. The Editorial Board were the Heads of the University of Wales's Welsh Departments. It is not surprising therefore that it was not felt that there was enough material to publish until 1948. The eagerly awaited first part entitled: *Geiriadur Prifysgol Cymru: (University of Wales Press) A Dictionary of the Welsh Language* appeared in 1950. The first volume was not completed till 1966 and the entire project will not be completed until 2002 - after nearly 80 years! Even then the team will take until 2008 to completely re-edit - letters A and B.

In the meantime other smaller dictionaries have appeared. One of these was the modest *New Welsh Dictionary* in 1953 which was revised and enlarged to include *a large number of technical words* - and entitled *Y Geiriadur Mawr* - in 1958. In this publication Welsh was 'a medium of thought and expression in the *modern world* and not only... a vehicle of a fine literary tradition'. It is interesting to note that the editors were two school masters, rather than University academics, their venture into lexicography coinciding with the beginnings (and rapid growth) of Welsh medium education. The first Welsh primary school was established in 1947 and by 1970 - 41 such

designated schools had been established. The first designated Welsh secondary school was established in 1956 and they too had expanded to 14 by the mid 1980's. There was therefore a desperate need for technical terms as more subjects were being taught and examined through the medium of Welsh. Today 20% of our school children are taught through Welsh.

From 1950 onwards the Board of Celtic Studies sought to face this problem by collating the terms used for grammar, phonetics, ethics, metaphysics, aesthetics, logic, music and chemistry in University classes and periodicals. Panels were also created by the Welsh Joint Education Committee to work on other fields - Education, Geography, Mathematics, Biology, Physics, Athletics, Games and Recreation, Cookery, Needlework, Embroidery, Knitting and Laundry work, Business and Office. Lists of scientific terms were created based on contributions to a Welsh Scientific Journal *Y Gwyddonydd*, other sources being radio and TV terminology, published in a magazine called *Arolwg.* In all about 100,000 items were available by 1973 to form the basis of a *Dictionary of Terms.*

The creation of the dictionary only served to highlight longstanding problems. The editors faced duplication of terms, inconsistencies, imprecision, etc., terms having been arbitrarily created without recourse to any criteria. The situation prevailed until the creation of a Centre for Terminological Standarization in the early nineties which my colleague Delyth Prys will talk to you about later on. Suffice it for me to say that the publication of the terminological dictionary: *Y Termiadur Ysgol*, of which she was editor, is an important milestone in the history of Welsh Lexicography.

Another important development - that occurred during the prolonged gestation period of the University of Wales dictionary - was the so called 'linguistic revolution'. This was when linguists became increasingly interested in linguistic fundamentals, in the very basic ingredients of individual languages. The belief was that teachers could then focus the learner's attention on their very 'core'. This was the background to the publication, in 1968, of a *Learner's Welsh - English dictionary*. It was based on a study of the language used by a sample of Welsh-speaking eight year olds, by R. Cyril Hughes, and on similar research conducted by international linguists elsewhere. It was, for example, heavily indebted to the research underpinning 'Francais Fondamental'. The dictionary contained 2,500 items selected 'according to principles of frequency, availability and range and graded for [use on] three levels of teaching' namely the primary, secondary and adult levels.

However the most complete conventional dictionary we have to date is the *Welsh Academy English-Welsh Dictionary* - published in 1995. It is the latest offering in a long utilitarian tradition. In the words of the preface:

> One does not need to be interested in lexicography to know that English is constantly changing and that it is continually adding to its vocabulary, with the result that its resources are virtually unparalleled. When major languages like German and French are borrowing words from English, it is no wonder that a neighbouring language like Welsh should do so. Unless Welsh can offer a means of communication adequate to compete with English in every sphere of life, its speakers will be under pressure to borrow more and more words from English and will end up speaking a patois with the feeling of linguistic inferiority...

The basic question posed by its authors - Bruce Griffiths and Dafydd Glyn Jones was - 'how does Welsh convey the meaning of an English word or phrase, drawing on its own resources'. It was based on the English/French half of the 1975 *Harrap English-French / French-English Dictionary,* and its subsequent revisions, and was sponsored by the Welsh Academy - established in 1959.

But although considered a *tour de force* the work's Achilles's heel is that, although it was compiled using word processing, its data was stored as text files rather than as an electronic data base. The wealth of knowledge gathered is therefore not easily exploited.

This brings into focus the *'computing revolution'*, although initially - only the computer's compositional facilities were fully appreciated. A computer was merely a super-efficient typewriter. The first tentative steps towards exploiting its electronic data base potential was taken in 1991 when the Welsh Language Board convened a meeting in Cardiff to advise them on developing Welsh Language software. The thinking was that:

> 'if users come to believe that the full range of computing facilities can only be made available... through the medium of some major world language then this will be another psychological nail in the coffin of minority languages'.

The committee decided to prioritise the development of a Welsh Language computer spelling and grammar checking programme that would eventually be known as 'CySill'. The early period was spent producing a bilingual base computer dictionary - 'CysGair', containing over 48 thousand entries - capable of dealing with the specific and peculiar characteristics of Welsh. These include, for example, the fact that the language has as many as 4 initial mutations the most prevalent of these being governed by 26 rules - as well as an additional list of 6 important exceptions. Welsh verbs are also inflected, even so-called 'regular' verbs not being as regular as the denotation would suggest. Even a 'regular' verb-noun like *gweithio* (as cited by the authors of the programme Nick Ellis, Cahill O'Dochartaigh, Bill Hicks, Menna Morgan and Nadine Laporte) can have as many as 37 different word forms. Welsh also has 7 ways of forming the plurals of nouns and there can be a good number of options even within a single rule. This means that apart from the most frequently recurring suffixes the plurals of nouns in the end have to be 'exhaustively listed'. Another headache was the language's circumflexed 'w' and 'y'. To cover all eventualities the original word list had to be expanded by a factor of 10.

However, despite such formidable obstacles the programme achieved a remarkable degree of success. One of its strengths is that it is able to recognise the mutated forms of a word so that the user can look it up in the dictionary in its unmutated form. Other achievements were that the programme can handle contractions occurring both before and after words, can provide a list of suggestions for misspelt words, and provide an user dictionary where words not occurring in the main dictionary can be stored. It can also cope with informal and formal registers - crucial if a spell-checker is to adequately reflect every day language use.

The period since the programme's launch (in 1993) has seen further developments although a number of taxing challenges still remain. One development has been the inclusion of technical terms - as they were being standardized - in the pro-

gramme's base dictionary and of word categories not included in the original cache. For example '- aid' suffixed words (in English - '-ful') - *gwydraid / glasaid* (glassful), *llwyaid* (spoonful), *sosbenaid* (saucepanful), etc. And recently, with the establishment of a National Place-names Centre and a newly constituted Place-names Advisory Committee, an opportunity to include 'standardized' place-names in the data base has become available. It is hoped, in the very near future, that Canolfan Bedwyr, the National Place-name Centre and Ordnance Survey will collaborate to produce a 'recommended' place-name list which will also be simultaneously made available to the 'CySill' data-base. The Centre is also inputting Welsh personal forenames and surnames, external geographical locations, the names of nations, official bodies and animals, birds, fish, flowers, trees, etc. into the basic dictionary.

Canolfan Bedwyr has also developed a Welsh Language hyphenator and is currently completing an electronic 'Thesaurus' for Microsoft Office.

A recently awarded project presents a further opportunity for data base expansion. Canolfan Bedwyr is currently engaged in converting the *Welsh Academy English/Welsh Dictionary* into an electronic data-base. Why are we doing it? Well, in the absence of a Welsh National Corpus this will provide us with valuable linguistic information, hitherto only enjoyed by the world's major languages. Together with the bilingual terminology databases it will, for example, highlight for us some of the language's *collocations* - the combination of words that occur so frequently that they are almost inseparable; 'When you see one you expect the other also to be there'. Some Welsh examples would be - noun+ adjective = *pobl gyffredin* (ordinary people), *tŷ bach* (toilet), *hel pres* (collection), etc. Another product could be a *bilingual idiomatic* dictionary able to highlight common or shared idioms and those that are not shared by the different languages. For example *talu* is used by both English and Welsh to 'pay' homage: *talu gwrogaeth.* We also 'sing' praises (*canu clodydd*) in both languages. But whilst we 'kill grass'-*lladd gwair* - in Welsh we 'cut grass' in English. But often the concepts concerned are much further apart. For example, Welsh uses a farming concept to denote the English 'show off' - *ceffyl blaen:* 'front horse' and the graphic *torri llengig* - the cutting of a muscle - for the much more subdued English 'rupture'. We also hope to produce a dictionary of *hyperonyms* and *hyponyms* and a 'better than nothing' word frequency dictionary. And we desperately need a dictionary of *Colloquial Welsh.*

The electronic databases will also be used as a resource - to facilitate 'gist translation', and improve our parallel text or memory translation capabilities (using such programmes as TRADOS). They will also pinpoint and quantify the most common word-building items. For example: an element like *budd* (originally part of the name of a successful Celtic Queen) has given us - *buddiol* (beneficial); *budd-daliad* (benefit payment); *buddugol* (victor/winner); *buddsoddi* (invest); *di-fudd* (useless/ unprofitable). Not only can identifying such productive elements help us to understand the language's present make-up but assist in the process of denotating emerging concepts. The dictionary will also, of course, be produced as a CD, and or in an online format.

As well as creating lexical databases Canolfan Bedwyr has started to create a Welsh Language electronic Corpus. It has accumulated a corpus of roughly a million words of written prose 'based on 500 samples of approximately 2000 words each,

selected from a representative range of text types... mainly post 1970'. The sample includes novels, short stories, children's factual and fictional writing, religious prose, non-fiction samples from the fields of science, education, business, leisure, local and national newspapers, public lectures, magazines, academic writing, biographies, etc. But, as the authors would readily concede, further work needs to be done to refine the categories and coding. For example, little attempt was made to refine the 'representative range of text types' by, for example, including *the different types* of factual writing - discursive, transactional, etc. The sample of 'academic' writing is again restricted. 'Children's writing' denotes **another** extensive group even if one merely thinks in terms of crude age ranges. In the survey, however, they are truncated into a single category. Again scant consideration was given to such issues as: translated texts, topical representation, the language's written and spoken forms and their fields of use or domains, the range of registers employed, etc. Neither are there any oral contributions in the compilation. A lack of resources also led to miscodings that in the opinion of the authors 'would require [at least] two years work' to rectify.

Canolfan Bedwyr has ambitious plans to build on the foundations laid. We are at present actively seeking funding to create a Welsh National Language Corpus. It will be on similar lines to the British National [English] Corpus though, of course, it will not be on anything approaching the same scale. The British National Corpus has 100 million entries -10 million of which are taken directly from spoken sources. This, when coupled with the American Corpus - to become the 'Bank of English' - becomes a staggering 400 million, of which 20 million is informally recorded (and transcribed) speech. The Welsh Corpus will obviously be more modest and will be based distinctly on studies of current Welsh usage. The obvious difference will be - much narrower fields of use. For example, the Welsh language is comparatively absent from further and higher education, whilst thriving in the primary and secondary fields. It is again relatively absent from the major fields of commerce and industry. There is no daily press yet it is extensively used daily on radio and on television. It is again spoken daily by half a million people. There is also a disparity of use between formal and informal situations. All this would have to be reflected in any representative corpus.

Why do we need a Corpus? In order to provide us with more reliable information - based on what people actually say in real situations. It will take cognizance of the actual use of present day Welsh. This will allow us to create dictionaries where the information has been checked against a 'large amount of corpus data' to give them 'reliability and authority'. It will enable us to 'make statements about ... meanings, patterns, and uses of words with much greater confidence and accuracy of detail'. It will enable us to replace made-up examples with examples of actual language that has been used. It will also allow us to discover things about our language that could never be gleaned, however hard we tried, by the human mind. Using computers opens up a whole new dimension of understanding. For example, with a Corpus we can begin to answer some of the following questions:

- which words, expressions and patterns are most commonly used?
- how accurate are our language definitions?
- which are the dominant meanings carried by a word?

Needless to say, this information will impact on grammars and teaching resources. Until now all our language 'descriptions' have been based on observation and hunches.

A Corpus will also bring into focus important issues. For example: is it possible to reconcile the purist view of language - that language ought to go through a filter before being recorded - with the perception of the lexicographer as a language recorder? [In Wales Sylvan Evans deliberated - long and hard - on every single entry; others (like Gweirydd ap Rhys) just put everything he could lay his hands on into the collection!] Can the views of those who feel that a language must be renewed from within itself - using its own resources - ever be reconciled with the 'internationalist' view seen at its sharpest perhaps when considering terminology? Should Corpus input be confined to those with an adequate command of language? Is there a threshold *of* language proficiency below which we should not go? (I myself believe in the validity of all attempts at expression). To what extent should an editor be aware of the conditioning influence of dictionaries, for example in schools? How do we reconcile the role of dictionaries as historical records, with their role as a living medium, the means by which we cope with the contemporary world? Lacking, as we do, the critical mass and technological vitality needed for natural language creation, what sort of intervention strategies should be put in place to ensure that our respective languages keep abreast of the developments of the times? These, and others, are momentous issues with which we will *all* have to contend.

Before handing over to Delyth, may I just say a few words about the MELIN project, because it is a project in which you and us were closely involved. The aim was to create multi-lingual - Basque, Catalan, Irish, Welsh - interactive, on-line, terminology dictionaries on the Web. Although the project did not reach its full potential it did at least succed in proving one thing namely that although there might not much common linguistic ground between the minority languages there is merit in producing a structure for common use. It also highlighted the constant need for minority languages, as well as sharing common experiences, to be *collectively* involved in mainstream technological and theoretical deliberations.