Fifth International Conference on Historical Lexicography and Lexicology
Oxford, 16–18 June, 2010

Second time around: re-editing *Geiriadur Prifysgol Cymru*

Andrew Hawke, *Geiriadur Prifysgol Cymru*, Aberystwyth

Geiriadur Prifysgol Cymru is the standard historical dictionary of the Welsh language. The project commenced in 1921 when a reading programme was established which it was intended would take five years to complete, with a further five years for editing the dictionary itself. In the event 27 years passed before work commenced on the editing, which began in 1947/8 with the first fascicle appearing in 1950. By this time it had been decided that a shorter version of the dictionary, more akin to the *Shorter Oxford English Dictionary*, would be published initially over ten years, to be followed by a fuller version in due course. Also, presumably because it was felt that the reading programme had been inadequate for the modern period, it was decided that no citations would be given after about 1800, although words and senses postdating 1800 would be defined as usual. A date noting the first attestation would be given, with no bibliographical reference. This considerably reduced the bulk of the Dictionary but made it far less useful for those readers who were primarily interested in contemporary Welsh.

Towards the end of the letter *B* it became increasingly apparent that funding for a separate fuller version would not be forthcoming, and gradually the scale of the dictionary was increased until it became much more akin to the OED in scope and appearance. In fact, the OED was instrumental in the planning of GPC.

Several members of the new GPC editorial board visited the OED offices around 1920 and received advice from Henry Bradley, one of the Co-Editors, then in his late 70s. Oxford University Press subsequently typeset and printed two of the four volumes and printed a number of fascicles of the third up until the sad demise of printing operations here in Oxford.

After 52 years' editing and 61 fascicles, the dictionary was finally completed in 2002 and launched at a ceremony in the Welsh Assembly in Cardiff. As soon as drafting the first edition reached the end of the alphabet in 2001, HEFCW (the funding council for higher education in Wales) and the University of Wales jointly agreed to fund the re-editing of the two unacceptably condensed initial letters of the alphabet.

* * * * *

It had been assumed that this would be a relatively straightforward task, mainly involving the addition of recent vocabulary (largely from the Dictionary's existing citation collection) and the incorporation of further citations to the existing entries (again, mostly from the existing collection). In fact, the scope of the revision has been far greater than originally envisaged, and in effect the dictionary has been completely re-edited rather than just being revised. The number of neologisms added has been far greater than had been anticipated, partly because of the increasing amount of material in Welsh available on the Web. The A–B section, has so far (up to *bin*) grown from 278 pages to 640, approx. 2.3 times the size containing 12,757 entries against the original 9,985, i.e. an additional 2,772 entries, an increase of 28%.

Lack of storage space has meant that none of the original drafts of the Dictionary have survived from the early days. However the original citation slips have survived although initially we were surprised at the lack of notes written on the slips themselves. Having the first edition as a basis for the new edition was, of course, a great help, but it also entailed certain drawbacks. Welsh (particularly poetic Welsh) includes many more simple compounds than English, and the meaning of most is self-evident from the component parts. These have often been excluded from the Dictionary. However, where the first edition editors included a word which would not meet our normal criteria for inclusion we have felt obliged to retain it in the second edition. Over the years the principle had been developed that three independent examples of any word were required before it would be considered for inclusion. This rule had been bent on occasions in the first edition when important, well-known, or influential hapaxes came under consideration, but we were surprised to see some fairly ordinary hapaxes included. These we have felt obliged to retain.

More of a problem were nonce-words or erroneous forms which had been included. In such cases we have redirected the reader to the correct entry where possible, or, failing that, have included a reason for excluding the word. We also encountered the problem of 'missing' citation slips. In the days before photocopying became generally available, it must have been tempting to move slips with good examples of other words on to another location in the alphabet, rather than copying them. As I mentioned before, when a post-1800 citation was the earliest attestation of a sense, just a date was provided. Where the slip had subsequently been moved to another location, we have no way of checking that citation, as we do not even know the author or title of the work. Rather than omitting such citations from the dictionary we have left them on the assumption that the earlier editors were correct, but have prefixed them with a question mark to warn our users that we cannot provide a further reference. In the first ten parts we have had to do this on 101 occasions.

A more serious problem occurs where the original editors were over-specific in differentiating various senses, perhaps under the influence of the corresponding word in the OED. When only a few citations were needed to illustrate each sense, it was of course much easier to find good examples and to ignore the dubious or ambiguous ones. We have found on many occasions, once all the citations for a given word are reconsidered, that we cannot sustain the original distribution of the various senses, and have been forced to conflate some of them to avoid misleading our users. It could be argued that we are, by doing this, reducing the usefulness of the second edition as compared with the first, but we have decided that this must be done in order to be academically rigorous.

Medieval Welsh poetry contains a large body of strict-metre verse, with very rigid rules for consonantal agreement, end-rhyme, internal rhyme, and syllable count. This severely restricted the poets' choice of words. Some also attempted to be deliberately archaic in their diction and some of the poetry is vaticinatory, and therefore deliberately ambiguous or abstruse. As in any language, the poets also played on the ambiguity of certain words to enhance their compositions. Figurative allusions were also extremely popular, which may be obscure to a modern reader. Add to this the fact that much poetry was originally preserved orally, or only survives in later (often multiple) copies based on several differing originals, and you can imagine the problems for the textual editor. All these things conspire to make some of the poetry extremely difficult for the lexicographer.

Where we have found it impossible to interpret the material in the way required to adhere to the first edition's sense divisions, we have often been compelled

to conflate sense paragraphs. In addition, very little of the medieval literature, especially the poetry, had been adequately edited when the first edition began, and much was cited from poor editions or directly from manuscript sources although they did certainly benefit from John Lloyd-Jones's magisterial *Geirfa Barddoniaeth Gynnar Gymraeg*, an excellent dictionary of the early poetry, which had been appearing in parts since 1931. They evidently used that work extensively, but also treated the earlier poetry less thoroughly because they assumed that their users would also have access to Lloyd-Jones's work. Unfortunately the *Geirfa* was left unfinished on Lloyd-Jones's death in 1956, although a final part up to the word *heilic* was published posthumously in 1963, and his widow generously gave the Dictionary staff access to his slips, which contain many valuable insights. By now the situation has improved considerably, and there are good modern editions of most of the important works.

At the time the first edition was commenced the Welsh language was in serious decline, with the number of speakers decreasing quite rapidly, and hardly any official recognition and even less support. Comparatively little was published in Welsh at that time, certainly much less than had been produced in the 19th century when there were over a million speakers. Because the target readership (of the first edition) of the Dictionary was assumed to be first-language Welsh speakers, Welsh was chosen as the metalanguage of the Dictionary, although English synonyms were given following the full Welsh definitions for each sense. The grammatical tags, notes on dialectal forms and place-names, and the etymological notes were all provided in Welsh only.

With the increasing number of Welsh learners, and the growing use of Welsh in commerce, and, especially in education and public administration, the user base has changed over the decades, until it is now much more mixed, and the Dictionary is used worldwide as the standard reference for academic study of the Welsh lexicon. In following closely the style of the first edition, the new edition has done little to alleviate this problem, and indeed to do so would have meant some fundamental changes. One thing that could be done, however, without affecting the basic plan of the Dictionary, was to provide more cross-references from certain forms that, whilst perfectly normal in terms of the language, create particular difficulties for learners or those completely unfamiliar with the language. In the first ten parts of the second edition some 2,588 additional cross-references have been added to the meagre 727 in the first edition.

The number of new cross-references is lower than appears, as the opportunity has been taken of disentangling multiple combined cross-references, which admittedly saved column inches, but proved difficult for users. This was also done with an eye on future electronic versions, where unambiguous cross-references will be necessary and it will not always be evident what is the order of the entries – e.g. a cross-reference may be required to the immediately following or preceding entry which would not be necessary in the printed version.

Another problem, exacerbated no doubt by the small size of the language community, is that of circularity. A scholar working on a particular text may have proposed a meaning for a word, which was then adopted by the Dictionary editors, who perhaps only had access to a few obscure citations. Once that meaning was published in the Dictionary, however tentatively, it was far more likely to be used by subsequent editors of other texts. We have come across a few examples of this whilst re-editing the Dictionary. Once we have a number of further citations of a word, a

quite different meaning may suggest itself which is more apposite than the original proposal.

Unfortunately the Dictionary may actually mislead scholars by misunderstanding the citations. One case in point is the term *betys arfor*, which is given plausibly in the first edition for 'sea beet'. This is in fact the result of a misleading citation slip which evidently had not been checked against the original text, where this term does not occur at all. All citations have been checked against the original printed or manuscript source for decades and we had assumed that this would have been the case in the early days of the Dictionary, however pressure of time may have precluded this. We spotted this particular mistake when we checked all the citations as part of our normal routine. By now, however, the term has been used in at least one modern English-Welsh dictionary and has been used to translate 'sea beet' on several occasions as a result. In such circumstances we have been forced to point out the error, and cite the first edition of our Dictionary as the first attestation! Another reason that this type of mistake is less likely today is that this was based on a single citation, whereas we would now only include a word or collocation if we had at least three independent sources.

By far the greatest change that has occurred is the availability of material on the Internet. The Dictionary was an early adopter of microcomputers, computerizing in the mid-1980s. This brought with it many benefits, gradually increasing as the technology improved and more and more Welsh-language material became available. However it was not until near the end of the first edition when the sheer volume of material in Welsh on the Internet first struck us. From that point onwards we have made increasing use of general Google searches.

Welsh is an inflected language, more akin to, say, German than English, but it also makes extensive use of initial consonant mutations, a feature of all the Celtic languages, and also internal vowel affection. This means that not only can the end of a word change, but also the beginning and the middle! As you can imagine, this makes electronic searching rather complicated, even using the modern standardized orthography. The computer notation known as 'regular expressions' is capable of dealing with this problem, and permits us to search texts stored locally. After many years of trying different products we eventually found the excellent program *FileLocator Pro* from Mythicsoft which does nearly everything we require. However, searching the Internet is another matter. We have to search multiple times for various inflected and mutated forms which is error-prone as well as frustrating.

The best solution is to download the Welsh-language material from the web and then search it locally. We have been very fortunate that Prof. Kevin Scannell of St Louis University offered to make us a web corpus, as he has done for many minority languages. His software analyses the language of a web page to determine whether it contains at least 50% Welsh material. If it does, it strips out the HTML tags, navigational aids and scripts, leaving plain text which is then saved on disc. Any hyperlinks on that page are then followed and the same procedure applied repeatedly until it runs out of Welsh-language material. The software has so far collected some 120,000,000 words of Welsh text. Kevin Scannell has produced a sorted list in Welsh alphabetical order together with the number of occurrences in the corpus. This enables us to know what words within the alphabetic range we are currently revising to search the Web for for further examples. Searching the Web for 'new' words can only be successful if you know what words to search **for**, of course!

By comparing the number of occurrences of various words in the corpus with Google searches for the same words, it appears that Kevin Scannell has succeeded in locating between a quarter and a third of the material accessible to Google.

Of course, we do have other evidence for most of the new words: in the citation slips that have accumulated since the publication of the first edition some sixty years ago, in electronic texts, lists of technical terms, other dictionaries, and so on. Online resources for Welsh have grown enormously over recent years:

- EEBO = Early 'English' Books Online,
- ECCO = Eighteenth Century Collections Online,
- EEBO/TCP = EEBO (Text Creation Partnership),
- ECCO/TCP = ECCO (Text Creation Partnership),
- 19th.-cent. Newspapers (BL),
- Cylchgronau Cymru (NLW) = Welsh (and Welsh-language) Journals,
- Dafydd ap Gwilym.net,
- Rhyddiaith Gymraeg, Middle Welsh prose texts
- Cambridge Historical Corpus of Welsh prose texts 1450–1850, etc., etc.

We also make extensive use of Google Book Search, which has made a large body of Welsh-language material available for searching (largely from US college libraries). This is more useful for the earlier material (generally late 18th- and 19th-century) because it is available in 'full view', permitting us to cite it freely, whereas the later material is often only available in 'snippet view', which is often insufficiently long. Annoyingly often, and for no apparent reason, some even older works cannot be viewed at all. Worse still is the metadata which accompanies some of these scans: dates are often erroneous, particularly in the case of journals. Often the only answer is to locate and check an original copy, which is, of course, very time-consuming, although we are fortunate in having a physical connection to the National Library of Wales, one of the six UK copyright libraries.

The major problem with Web material (as opposed to the major online resources) is that it is so transient and unattributable. Most online Welsh material is either of an official or commercial nature (and usually the work of professional translators), or else is from colloquial or informal sources such as blogs and online discussion forums. Whilst the latter can provide useful colloquial citations, the material is usually very difficult to attribute and to date, and almost impossible to refer to (certainly in any permanent way). In the future, this can perhaps be solved by permanently saving snippets of texts, with their last known location, and the date they were accessed. These snippets could then be linked to from the online dictionary.

A further annoyance is automatically translated material (either by Google Translate or an eastern European system that simply translates on a word-for-word basis with no parsing, generating Welsh-looking gibberish). Because Google Translate in based on parallel texts, the standard of its translations can be very high - high enough to fool web crawling software. The high frequency of certain improbable phrases generated by the other system means that its output can usually be stopped from entering the web corpus. The better these translation systems become, the greater the problem for future lexicographers, particularly for non-English languages.

Altogether the total number of citations has grown from nearly 28,600 to over 68,000, an increase of nearly 40,000. Similarly, the number of collocations has risen from 1,250 to over 1,900. The number of 19th-century citations has risen from just 53 to over 2,500, and 20th-century citations are up from under 20 to about 450, still a

comparatively small number. This is partly offset by the number of plain dates given for the first attestations of particular senses, which have gone up dramatically for the 20th century from about 360 to nearly 2,200. As new items are added to our bibliography, enabling us to cite from them fully in the Dictionary, these should be replaced by full citations. Citations from the present century are surprisingly rare: only 28 have been added so far, perhaps reflecting the difficulties of collecting the most recent neologisms.

As part of the general revision, we are also reviewing all the etymologies in the Dictionary. There have been considerable advances in Indo-European and Celtic etymology, including laryngeal theory. Again we have been very fortunate in securing the help of two eminent Indo-Europeanists, Peter Schrijver and Stefan Schumacher, who both have extensive experience with British Celtic. Some 3,086 etymological or morphological notes have been added to the existing 10,012 in the first ten fascicles.

As we approach the end of B, we are planning the next stage of the work. The intention is to migrate our present data to a new dictionary writing system and a modern corpus query system. Following a consultation with some of our users, it has been accepted that the Dictionary will henceforth only be published online in full, replacing the concise PDF-based version currently available via our website.

We are currently conducting a pilot project to assess the number of completely new words that need adding throughout the alphabet, in the hope that we may be able to edit this material quite rapidly and add it to the online Dictionary to bring it more up to date. Our experience of re-editing A-B has demonstrated that there is a pressing need to revise the whole work on a continuous basis, whilst being freed at last from the appropriately named ‘tyranny of the alphabet’ will permit us for the first time to update the Dictionary and correct errors where they are most needed.

It will be rather sad to see the end of the Dictionary as a printed work after so many years, but the OED has shown that such an approach can be very successful and it is obvious that the benefits far outweigh the disadvantages.