Quantifying the Use of Digital Welsh-language Language Resources

Delyth Prys, Gruffudd Prys, Dewi Bryn Jones

Bangor University
College Road, Bangor, Gwynedd, Wales
{d.prys, g.prys, d.b.jones}@bangor.ac.uk

Abstract

This paper quantifies the use of a number of Welsh-language language resources, namely electronic dictionaries, corpora and proofing tools developed at one institution in Wales, in an attempt to measure and compare their uptake by the language community. The number of searches and downloads of some of these resources over a number of years are tracked, with possible reasons given for certain trends. This will provide a baseline for further research and more sophisticated analysis of the results, helping to plan for improved dissemination and marketing, engaging with all stakeholders, including developers who are tasked with developing new content, platforms and media.

Keywords: LRL, dictionaries, terminology, proofing tools, corpora, digital media, apps, Welsh

1. Introduction

Welsh is a Celtic language spoken by approximately 562,000 speakers, or 19% of the population of Wales (Census Returns 2011). It suffered steep decline during most of the twentieth century, but its position has stabilized somewhat during the first years of the twenty first century, helped by the growth of Welsh-medium education and campaigns for revitalization. The development of digital resources and language tools are seen as a crucial part of the strategy to ensure that it survives and flourishes in the new millennium.

This was confirmed in the Welsh Government's current five year Welsh Language Strategy, *A living language: a language for living* (2012). The sixth strategic area in this document is Infrastructure, with the desired outcome of "More tools and resources in place to facilitate the use of Welsh, including in the digital environment". This was further elaborated as including terminology standardization, lexical resources and e-publishing. Crucially it also included research and data amongst its priorities, stating that "in order to test the effectiveness of this work [increase the use of Welsh], we need baseline data on language use, and regular data collection to allow us to monitor progress against the desired outcome".

The aim of this paper therefore is to present and analyse statistical data relating to the use of a number of digital Welsh-language resources to which the authors have access. In addition to quantifying the number of users recorded as having used each resource, the paper will examine the frequency, time and origin of 'usages' so that usage patterns of significance may be identified.

Where applicable, the statistics for each resource are compared with those of the other digital Welsh-language resources included in the exercise. Due to the practical considerations of requiring access to such data and the permission to publish said data this paper concentrates exclusively on resources developed at Bangor University.

These resources include websites, web-based linguistic services and apps, all of which are lexical or terminological in nature. They are presented individually by category, and are accompanied with a brief description of the resource to provide context to the bare statistics which are reported as they were as of 18/09/2015. A table featuring the collated statistics for all the individual resources follows the presentation of individual resources. Comparisons that can be made between the individual resources and categories are explored in the paper's conclusion.

2. Websites

The websites can be divided into lexical, terminological, hybrid lexical/terminological and proofing resources.

Google Analytics is installed on these websites to enable the monitoring of user traffic to the resources provided (many of the funders of these resources require the reporting of such figures as part of the agreement to finance the projects that create and maintain these resources).

Searches and other forms of text submission however are logged in databases on projects' own servers and constitute a more fine-grained measurement of the use of resources as a single visit to a resource's webpage can result in several unique searches.

2.1 Websites: Lexical Resources

2.1.1 Geiriadur Bangor

Geiriadur Bangor (The Bangor Dictionary) (2011), is a searchable online bilingual bidirectional Welsh/English dictionary which replaced the BBC's retired Welsh/English *Learn Welsh Dictionary* (n.d.) which had been primarily aimed at learners of Welsh as a second language. It combines a general language dictionary, *Cysgair2* (2004) and *Y Termiadur* (Terminology Dictionary) (Prys et al 2006), an older version of *Y Termiadur Addysg* (Terminology Dictionary for Education) (Prys and Prys 2011), the standardized terminology dictionary developed for bilingual primary, secondary and further education in Wales. The dictionary entries include the source language headword, a disambiguating part of speech (to differentiate between the noun and verbs in English that share the same word form), concept disambiguation text (to differentiate between concepts that share the same word form), as well

as the equivalent word form(s) in the target language, their parts of speech and their plural forms.
Launch Date: October 2014
Total Searches: 97,872
Avg. Searches/Month: 4,255

2.1.2 Geiriadur yr Academi - The Welsh Academy English-Welsh Dictionary Online

Originally developed for the Welsh Language Board (now replaced by the Welsh language Commissioner) this online version is a digitized, freely accessible version of *Geiriadur yr Academi - The Welsh Academy English-Welsh Dictionary* (Griffiths and Jones 1995) first published in print form and later digitized to produce an on-line version. Based on Harrap's English-French dictionary it is considered to be the first modern comprehensive unidirectional English to Welsh dictionary and serendipitously appeared in time to help address the increased translation needs of post-devolution Wales. Enhanced search facilities and free access has improved access to the dictionary's contents. In addition to headwords, parts of speech, target language equivalents and their plurals, the dictionary includes phrases and their translations.
Launch Date: February 2012
Total Searches: 5,048,855
Avg. Searches/Month: 117, 415

2.2 Websites: Terminological Resources

2.2.1 Y Termiadur Addysg

Y Termiadur Addysg is an online terminology dictionary funded by the Welsh Government to provide standardized terminology for pre-university Welsh-medium and bilingual education in Wales. Welsh-medium educational materials such as exams, assessments and course materials are required to use the terms published in the *Y Termiadur Addysg* dictionary. The dictionary is therefore targeted at teachers, students and their parents, as well as those engaged in production of Welsh-language educational materials, including both original and translated works. Its entries are also included in the *Geiriadur Bangor* dictionary cited in 2.1.1 above.
Launch Date: August 2011
Total Searches: 568,136
Avg. Searches/Month: 11,595

2.2.2 Geiriadur Termau'r Coleg Cymraeg Cenedlaethol

Geiriadur Termau'r Coleg Cymraeg Cenedlaethol (Andrews and Prys 2015) is an online bilingual Welsh-English terminology dictionary for Higher Education. It began to be published online as separate dictionaries for each subject field in 2010, and was merged into one dictionary in July 2015. It currently covers 14 subjects taught through the medium of Welsh at university level. It deals with a narrower range of academic subjects than *Y Termiadur Addysg* but does so in greater depth. It also includes definitions, illustrations and diagrams.

Launch Date: March 2010
Total Searches: 18,824
Avg. Searches/Month: 285

2.2.3 Welsh National Terminology Portal

The Welsh National Terminology Portal (Jones et al 2011) is an on-line dictionary portal that aggregates the content of 18 bilingual Welsh-English terminology dictionaries, allowing them all to be searched using a single search query. The aggregated dictionaries cover a variety of domains including education, health, social care, justice and natural science, and include *Geiriadur Termau'r Coleg Cymraeg Cenedlaethol* and *Y Termiadur Addysg*, described above. The majority of the earlier dictionaries that are included have been published in print form in addition to their online form, though increasingly they are now only published online.
Launch Date: March 2010
Total Searches: 836,414
Avg. Searches/Month: 12,673

2.3 Websites: Proofing Resources

Proofing resources differ from dictionary searches in that they receive free text submissions rather than searches in the form of words or multi-word terms or entities. In submitting texts and utilising the proofing interface, the users of online proofing tools provide data in the form of both the texts to be proofed and the proofing decisions made

2.3.1 Cysill Ar-lein

Cysill Ar-lein (2009) is a free online tool for Welsh grammar and spell checking. Based on the commercial *Cysill* grammar and spell checking programme, part of the *Cysgliad* (2004) software package for Windows, it is limited to checking texts with a length of approximately 3000 characters. As stated in the Terms of Use, submitted texts are stored as a corpus for academic research purposes and are a useful source of word frequency counts, neologisms, common errors, dialectal forms and named entities. The corpus is not currently publicly available due to issues of confidentiality.

The texts passed through Cysill for spelling and grammar checking is varied, including school and student essays, journalist and scholarly articles, blog posts and tweets, and even job applications. It seems therefore to be used by a broad range of people, from children upwards, and for a broad variety of purposes.
Launch Date: August 2009
Total Submissions: 1,695,912
Total Words Submitted: 31,114,203
Avg. Submissions/Month: 23,231

2.4 Websites: Corpus Resources

2.4.1 Welsh National Corpora Portal

The *Welsh National Corpora Portal* (2011) website hosts a number of Welsh-language text corpora. At present these include two 'editions' of the English-Welsh parallel corpus *Proceedings of the National Assembly for Wales* (19.5 million words), the *CEG Electronic Corpus of*

Welsh (1 million words, monolingual), the *DECHE Corpus of Welsh Scholarly Writing* (currently 334,000 words, monolingual), and the *Example Corpus of Language Registers* (1,800 words, derived from the unpublished 22+ million word monolingual *Cysill Ar-lein* corpus). They are all searchable online and simple to use.
Launch Date: April 2011
Total Searches: 307,469
Avg. Searches/Month: 5,801

3. Web-based Services

The following resource is a web-based service. As such it is active primarily on websites hosted by external institutions and therefore information regarding the number of site visits was not available for this paper.

3.1 Vocab Bangor

Vocab Bangor (2015) is a website widget that enables users of websites utilizing the service to access English translations of Welsh words by hovering the cursor over the word in situ. This enables learners or the readers of very technical texts to better understand the content of the website without recourse to external dictionaries. Currently, the most high profile use of the service is the Welsh-language daily news website www.golwg360.cymru. *Vocab Bangor* replaces a similar service retired by the BBC.

In the statistics reported below, a 'search' represents a request for a translation equivalent in the form of a hover event for a word found in the text of the webpage on which the service is active.
Launch Date: March 2015
Total Searches: 30,020
Avg. Searches/Month: 5,003

4. Comparison of Web Resources

The table below gives an overview of the statistics for all the web-based resources.

	Total Searches/ Submissions	No. Months	Avg. Searches/ Submissions per month	Launched
GA	5,048,855	43	117,415	Feb 2012
GB	97872	23	4,255	Oct 2014
TA	568136	49	11,595	August 2011
PT	836414	66	12,673	March 2010
GTCCC	18824	66	285	March 2010
CYS	1,695,912	73	23,231	August 2009
CORP	307,469	53	5,801	April 2011
VOCAB	30020	6	5,003	March 2015

Table 1. Collated statistics for the web-based resources.

Key: GA: Geiriadur yr Academi; GB: Geiriadur Bangor; TA: Y Termiadur Addysg, GTCC: Geiriadur Termau'r Coleg Cymraeg; CYS: Cysill Ar-lein; CORP: Welsh National Terminology Portal; VOCAB: Vocab Bangor.

Direct comparisons are not possible, as some have been in existence for longer than others. Some are also targeted at different audiences, some general, some educational, and some for second language learners of Welsh.

5. Additional Aspects

5.1 Change in user numbers over time.

Figure 1 displays the total monthly searches during March for years 2010-2015 in order to gauge the growth of some of the more popular resources: *Geiriadur yr Academi* (GA), *Y Termiadur Addysg* (TA) and the *Welsh National Terminology Portal* (PT). March was selected as a fairly typical month for a month-based scale.

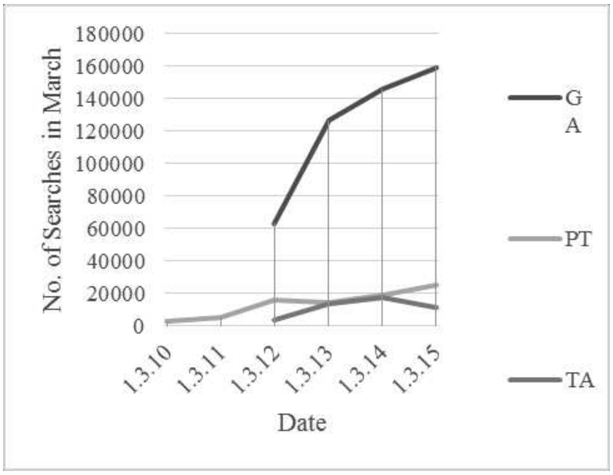

Fig 1. Monthly searches during March 2010-2015.

The fall in the number of searches on the *Y Termiadur Addysg* website can be explained by the realisation amongst users that its contents (and that of many more dictionaries) are incorporated within the *Welsh National Terminology Portal*, which has enjoyed a corresponding rise in figures. In general, the figures display a solid growth in the number of searches across all resources apart form *Y Termiadur Addysg*.

5.2 Weekly and Seasonal Variation

One feature of the statistics not reported in these statistics but apparent in figures 2 and 3 below is the consistent variation in search numbers on weekends and during the holiday periods over Christmas and during summer holidays. This suggests that heavy use is made of these resources during the working week, within both education and office environments. Much less use is made of them at weekends and during holidays, suggesting that they are not seen as resources for leisure use.

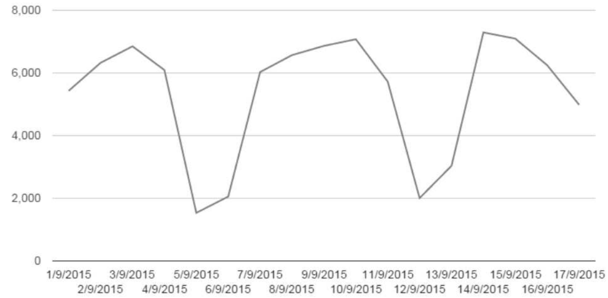

Figure 2. *Geiriadur yr Academi* searches during September 2015 showing prominent dips at weekends.

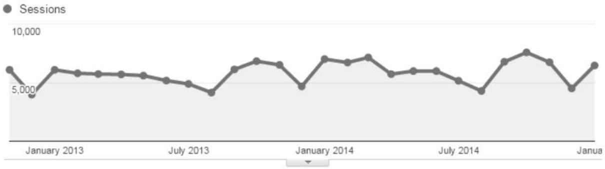

Figure 3. *Welsh National Terminology Portal* searches between December 2013 and January 2015 showing dips during holiday periods.

6. Apps

6.1 Ap Geiriaduron

The *Ap Geiriaduron* (2013) dictionaries app is a free searchable dictionary app containing the *Cysgair* general language dictionary, and *Y Termiadur Addysg* terminology dictionary for primary, secondary and further education. Following a recent update, it also includes G*eiriadur Termau'r Coleg Cymraeg Cenedlaethol*.

In addition to displaying general language information such as headwords, parts of speech, equivalents and plurals, it features entries with rich definitions including photographic images, diagrams and MathJax-powered mathematical formulas. The 'dictionaries' in the app's title is significant as the app has been designed as a medium into which further dictionaries may be added.

The *Ap Geiriaduron* has been well received by users, having an average rating of over 4 stars in each of the different App Stores.

6.1.1 Ap Geiriaduron - Total Downloads

According to *appfigures.com*, the download aggregating services used by the institution, *Ap Geiriaduron* has been downloaded a total of **46,191** times. This is a significant number in the context of language with a reported 562,000 speakers, and when compared to the average sales of traditional Welsh-language books.

6.1.2 Ap Geiriaduron - Downloads by Operating System

The *Ap Geiriaduron* is available for the iOS, Android and Amazon mobile device operating systems. Figures from *appfigures.com* indicate that the iOS version of the app has been downloaded in significantly greater numbers than the Android versions. The Android-based Amazon version represents a small proportion of the total downloads, but 1,208 is a fairly significant number in the context of less-resourced languages.

Operating System	Downloads
iOS	43,114
Android	13,184
Amazon	1,208
TOTAL	**46,191**

Table 2. Downloads by Operating System

6.1.3 Ap Geiriaduron – Cumulative Downloads over Time

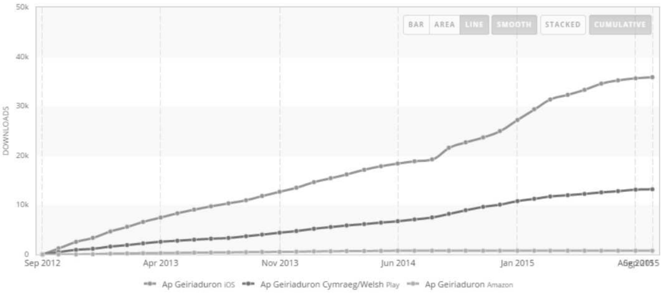

Table 3. Downloads over time

6.1.4 Ap Geiriaduron – Downloads by Country

It was also possible to track downloads according to the geographical location of downloaders. According to the table below, users were overwhelmingly from the United Kingdom, with relatively modest use in overseas countries.

Country	Style
United Kingdom	43,114
United States	1,181
Australia	260
Germany	139
Canada	129
China	63
Ireland	56
France	52
New Zealand	51
India	51
Argentina	42

Table 4. Downloads by Country (Top 11)

7. Conclusions

Dictionaries, corpora and proofing tools are essential aids for less resourced languages. Although it is difficult to obtain figures for previous sales of traditional paper versions, where they existed, print runs were never very large, ranging from 500 to 30,000 copies for bestsellers over a product's lifetime. Making them freely and easily accessible in digital formats improves their accessibility, with the figure of over 46,000 downloads so far for the *Ap Geiriaduron* being a particularly encouraging statistic.

The actual usage statistics for all these products is also encouraging, as anecdotal evidence suggests that paper dictionaries were previously bought but not necessarily consulted on a regular basis due to the dictionary not being at hand when required or the inconvenience of searching a paper dictionary. This may have been especially true of a very large and complex dictionary such as *Geiriadur yr Academi*, where some entries could be many pages in length. Searching the online version for an idiom buried in the middle of a large entry is much easier than searching the condensed format of the print version, and the 5 million online searches since its launch in 2012 testifies to its popularity.

The terminology dictionaries are more specialist in their content and audiences. They are also available in various guises, both federated and aggregated, making it difficult to count the popularity of any one terminology dictionary. Anecdotal evidence again suggests however that users value being able to search all terminology dictionaries through one portal rather than having to search dictionaries separately. This may account in particular for the lower figures for direct access to *Geiriadur Termau'r Coleg Cymraeg Cenedlaethol,* the terminology dictionary for Higher Education, as students will already be familiar with *Y Termiadur Addysg* from their schooldays, and wish to continue using it, or else use the National Terminology Portal to access all terminology dictionaries at once.

It is notable however that all these traditional forms of accessing on-line resources are being somewhat eclipsed by the popularity of the downloadable *Ap Geiriaduron*. Its popularity continues to grow, mostly through word of mouth, and it does not seem to have reached a ceiling yet in terms of number of downloads. The level at which it finally plateaus would be informative in terms of the active numbers of Welsh speakers and learners. Unfortunately, privacy issues do not allow us to gather background information on users, but the developers are exploring other means of obtaining information through voluntary feedback. The portability of the app, and the way it can be accessed in the classroom, the pub, when travelling and so on, make it eminently suitable to users for whom their mobile devices are an integral part of their lives. It also helps provide a minority language gain a more contemporary image in keeping with its desire to flourish in the 21st century.

Geiriadur Bangor and *Vocab Bangor* are more recent products, aimed primarily at Welsh learners who represent an important target audience.

While the popularity of the free, online version of the *Cysill* spelling and grammar checker was well-known, the relative popularity of the corpus resources in the *Welsh National Corpora Portal* came as a surprise. With over 5,000 searches per month, this was previously regarded as likely to be of interest only to researchers and academics. It appears however that translators consider it a good source of translation solutions as it contains words and phrases in context, a feature not often found in dictionaries.

All these resources therefore seem to be well received and used by their intended audiences, especially considering the lack of finance for dedicated marketing campaigns. There is clear evidence of growth in the use of mobile versions of the resources, and this will inform the prioritisation of future developments. These figures may also be of interest to language planners and policy makers in other minority language communities, as an indication of the digital resources likely to be most popular if developed.

However, this exercise has also highlighted a division between use for work and educational on the one hand, and recreational and leisure use on the other. While utilitarian language tools and resources are an urgent need for any less-resourced languages, users also need fun apps and resources for their free time. Both needs are equally important and should be catered for. It would be interesting to measure the use of any such apps for Welsh and other less-resourced languages, and compare their uptake to that of the ones presented in this paper.

References

A living language: a language for living (2012). Welsh Government, Cardiff.

Andrews, Tegau and Prys, Delyth (2015) *Geiriadur Termau'r Coleg Cymraeg Cenedlaethol*. Coleg Cymraeg Cenedlaethol: Cardiff.

Ap Geiriaduron (2013). Bangor University: Bangor. http://techiaith.cymru/widgets/vocab/?lang=en. Accessed 17/09/2015.

Census Returns (2011). https://statswales.wales.gov.uk/Catalogue/Welsh-Language/WelshSpeakers-by-LocalAuthority-Gender-DetailedAgeGroups-2011Census. Accessed 17/09/2015.

Cysgliad (2004). Bangor University: Bangor.

Cysill ar-lein (2009). Bangor University: Bangor. http://www.cysgliad.com/cysill/arlein/. Accessed 17/09/2015.

Geiriadur Bangor (2011) http://geiriadur.bangor.ac.uk. Accessed 17/09/2015.

Griffiths, Bruce and Jones, Dafydd Glyn (1995) *The Welsh Academy English-Welsh Dictionary.* University of Wales Press: Cardiff. Online version at http://geiriaduracademi.org/

Jones, D.B.; Prys, Delyth and Prys, Gruffudd (2011). *Welsh National Terminology Portal*: Bangor University, Bangor.

Jones, D.B.; Prys, Delyth a Prys, Gruffudd (2013) *Welsh National Corpora Portal*. Bangor: Bangor University.

Prys, Delyth; Jones, JPM; Davies, Owain and Prys, Gruffudd (2006) *Y Termiadur.* ACCAC: Cardiff:

Learn Welsh Dictionary (n.d.) BBC Wales: Cardiff. http://www.bbc.co.uk/cymru/geiriadur/. Accessed 17/09/2015.

Prys, Delyth and Prys, Gruffudd (2011-) *Y Termiadur Addysg*. Cardiff: Welsh Government.

Vocab Bangor (2015). Bangor University: Bangor. http://techiaith.cymru/widgets/vocab/?lang=en Accessed 18/09/2015.