Cysill Ar-lein: A Corpus of Written Contemporary Welsh Compiled from an On-line Spelling and Grammar Checker

Delyth Prys, Gruffudd Prys, Dewi Bryn Jones
Language Technologies Unit, Bangor University, Bangor, Wales, UK
E-mail: {d.prys, g.prys, d.b.jones}@bangor.ac.uk

Abstract

This paper describes the use of a free, on-line language spelling and grammar checking aid as a vehicle for the collection of a significant (31 million words and rising) corpus of text for academic research in the context of less resourced languages where such data in sufficient quantities are often unavailable. It describes two versions of the corpus: the texts as submitted, prior to the correction process, and the texts following the user's incorporation of any suggested changes. An overview of the corpus' contents is given and an analysis of use including usage statistics is also provided. Issues surrounding privacy and the anonymization of data are explored as is the data's potential use for linguistic analysis, lexical research and language modelling. The method used for gathering this corpus is believed to be unique, and is a valuable addition to corpus studies in a minority language.

Keywords: corpus, proofing tools, Welsh, minority language

1. Background

The *Cysill Ar-lein Corpus* is a written corpus of contemporary Welsh created through the novel means of collecting texts from an online spelling and grammar checking service, *Cysill Ar-lein*. Launched in 2009, *Cysill Ar-lein* is an online implementation of the popular *Cysill* Welsh-language spelling and grammar checking tool for Microsoft Windows, first produced in the early 1990s (Hicks, 2004) and updated, repackaged with a compendium of electronic dictionaries, and released as *Cysgliad* in 2004. Independent studies have since highlighted the popularity of this proofing tool and its role in supporting use of Welsh in computing, online and multimodal environments (Prys, 2008). In 2015, the *Cysill Ar-lein* interface was made available to external developers to embed in their own websites using an API key available from the *Welsh National Language Technologies Portal* (2015).

Figure 1: The Cysill Ar-lein interface

Whilst the original Windows version of *Cysill* remains part of the commercial *Cysgliad* package of Welsh-language tools and resources, *Cysill Ar-lein* was released as a free-to-use service. This was done primarily to enable the service's developers to gather as large as possible a corpus of contemporary written Welsh from user submissions. A secondary consideration was to ensure that a version of the software was available to all, regardless of their budget or operating system of choice, whilst promoting rather than undermining the sales of the commercial package which funds the continued support and development of the software. For these reasons, unlike the Windows version, the online *Cysill Ar-lein* service is limited to processing approximately 3000 characters at a time, and its users must accept terms of use that explicitly allow any submitted texts to be used for research purposes. All versions of *Cysill* do however make use of the same lexicon, part of speech tagger and grammar rules.

2. The Cysill Spelling and Grammar Checking Process

Cysill's spelling and grammar checking process will now be described briefly as a means of informing the discussion of the resulting corpus. Both the desktop and online versions of Cysill follow a near identical process when checking users' texts. The text is first checked for spelling errors, with any word forms not in the *Cysill* lexicon being identified as potential misspellings, and similar forms from the lexicon being suggested as possible corrections. For example, the following short sentence contains two errors, the first being that 'cath' (English: cat) is misspelled, whilst the second is a grammatical error:

Mae caht mawr yma.
English: There is a large cat[misspelled] here.

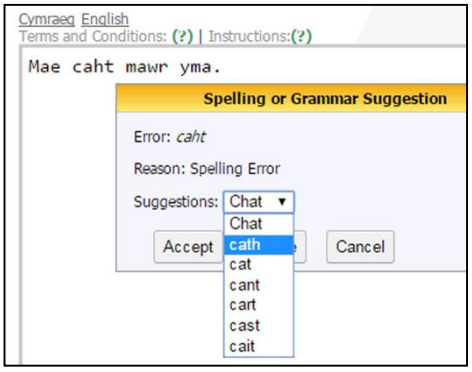

Figure 2: Spelling Suggestions

In this case, Cysill would first identify 'caht' as a possible spelling error and then display a number of possible correct forms for the user to accept or decline as substitutions. Only once the program has run through the entire text identifying potential misspellings and the user has had an opportunity to correct any perceived errors does the grammatical analysis begin. At this point a Welsh-language part of speech tagger is then used to classify the tokens so that a series of grammatical rules numbering in the hundreds may be run on the texts. If any of these rules are broken whilst running through the text, the user is notified with a message stating the expected grammatical behaviour and suggesting a correction.

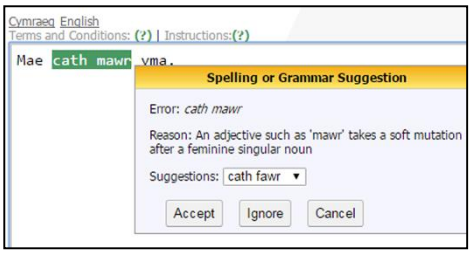

Figure 3: Grammar Suggestions

For example, whilst the following sentence no longer contains any misspelled forms, it lacks the initial consonant mutation required of an adjective that follows singular noun such as 'cath':

> Mae cath mawr yma.

Cysill will identify this failing and display a message explaining the problem, in this case:

> An adjective such as 'mawr' takes a soft mutation after a feminine singular noun.
>
> Suggestion: cath fawr

If the user accepts the suggestion by clicking on the 'Correct' button, the text is changed in place to:

> Mae cath fawr yma.

Once the suggestion has been implemented or declined, the program moves on to the next possible error that has been identified, should one remain. In some cases, such as with the following text, the user is presented with a number of options:

> sbectol haul dywyll
> English: dark sun glasses

The message displayed by *Cysill* is as follows:

> Do not mutate an adjective such as 'haul' after masculine and plural nouns.
>
> But if dywyll refers to 'sbectol haul' and not just 'haul' then you don't need a mutation.
>
> Suggestion: sbectol haul tywyll

In this case, it would be wise to ignore *Cysill*'s suggestion as the adjective 'dywyll' (dark) does in fact refer to the 'sun glasses' rather than the 'sun'. These examples have served to both explain how the submitted texts are processed by Cysill and highlight an issue with *Cysill*'s text proofing process, which is that these messages require a certain level of linguistic understanding to be understood and acted upon appropriately, something that anecdotal evidence from language improvement teachers suggest can be taken for granted by its users. The ability to analyse users' behaviour in such cases was one of the many aims in mind when establishing the *Cysill Ar-lein* as a means of collecting a corpus of user submitted texts.

3. Corpus Description

The Cysill Ar-lein corpus is a monitor corpus, i.e. it continues to grow as texts are added to it on a daily basis as users make use of Cysill Ar-lein to check the spelling and grammar of their written documents. When a user uses the service, a copy of the user's text is recorded at each point during the spelling and grammar checking process described above. These versions include not only the initial submission and the final proofed version of the text, but also a version for each spelling or grammar issue identified by the programme, along with the edits (or lack of edits) made by the user in response to the software's recommendations. This in effect generates two main corpora: an unproofread corpus and a proofread corpus, along with a difference trail of the user's responses to the software's recommendations that is useful for evaluating issues such as whether users understand and react appropriately to the software's recommendations, or refuse some of the recommendations as being inaccurate, inappropriate or incompatible with the document's register. Although the uncorrected corpus may be considered a 'corpus of errors', a manual analysis of a subset of the corpus undertaken in 2010 for an MRes project (Wooldridge, 2011) suggests that texts cover a wide range of subjects. Her analysis identified five categories of texts: Correspondence, Creative Writing, Academic Work, Publicity and Other. Correspondence was further subdivided as letters, emails and Twitter messages, Creative Writing as short stories, poetry and blogs, Academic Work as schoolwork and college essays, Publicity as adverts and news stories, and Other as Application forms/CVs, reports and fragments. No attempt was made to estimate the different percentage of each category of text. However, more recent manual reading of parts of the corpus confirm the presence of each category, and that they are easily identifiable due to their subject matter. Wooldridge surmises also that its contributors include a wide cross-section of users, ranging from proficient writers of Welsh, who use the online checker as a convenient tool to check for typing and other minor errors in their work, to second language learners with a very basic knowledge of the language. Requiring users to register their details would likely result in fewer users using the resource and increase privacy concerns, therefore the Cysill Ar-lein website does not collect information about the background of its users. This means that there is no extrinsic means of knowing the age, gender, language proficiency or similar details for the author of any text, which might be useful for some types of error analysis. However, text input by users is generally of sufficient length to allow researchers to

identify the user's language level, and the subject matter may also often be deduced from the context.

Some of the texts found within the corpus are of a sensitive nature. For example, job applications containing the applicant's name and address details, reports on individuals and other confidential documents have all been found within the corpus, and although a basic Welsh-specific anonymizer has been developed for use with the corpus, it has not yet been established whether its use would be adequate to protect people's anonymity without additional manual filtering. As a result, the *Cysill Ar-lein* corpus has not yet been published for general use, and remains a tool for internal research or a source from which smaller manually anonymized corpora can be produced, for example the *Example Corpus of Registers* (2014).

4. Analysis of Use

As of September 15th 2015, the *Cysill Ar-lein Corpus* contained over 31 million words from a total of 1.7 million submissions. However, this figure represents the raw input from the website and does not account for duplicated submissions or the surprising amount of English-language submissions. Nevertheless, this represents one of the largest digital corpora of Welsh to date. The popularity of the website has grown steadily since its inception, with reductions during school holidays in Wooldridge's analysis from 2009 indicating extensive use in schools (Wooldridge, 2011). By September 2015 date there were, on average, over 600 daily users of the website, generating approximately 1,100 daily page views.

A detailed analysis of use was made during the 11 months between August 2009 and July 2010, tracking the increase in use and the types of text corrected using the on-line checker (Wooldridge, 2011). Google Analytics was installed on the website in June 2010 in order to allow researchers to further track the popularity and use of the on-line service. Statistics reported by this method are extremely high, but the more modest statistics collated from the internal database itself still show significant use of Cysill Ar-lein. During the period between 11/03/13 and 09/03/15, according to the internal database statistics, there were 79,723 unique users (counted as number of different computers used to access the site), giving a total number of 1,706,245 sessions where text was input for automatic proofreading by the system. In some cases the text could have been a single word or a very short piece of text, but given the upper limit of 3, 000 characters (around 500 to 600 words), no text input could have been longer than this. Anecdotal evidence, in the form of numerous requests to raise the upper character limit from users currently processing large documents in smaller chunks, suggest that many texts are at the upper limit of the range. Average time spent on the website during this period was 7.13 minutes, indicating that users were spending considerable time inputting texts, considering the suggested corrections, and either accepting or rejecting them.

Out of a total population of 562,000 Welsh speakers (2011 Census, Key Statistics for Wales, 2011) these figures are surprisingly high, especially in view of the fact that the *Cysgliad* commercial product has also been available during the same period, as well as free Welsh language spellcheckers from Microsoft Word and OpenOffice. Anecdotal evidence suggests that this may be due to the lack of confidence speakers of Welsh, as a minority language, may have in their ability to write the language, especially as the older generation would not have had any education through the language, and it is still under-represented in official domains.

Whatever the reason, it has provided researchers with a valuable large-scale corpus at very little additional cost or effort.

5. Corpus Tools

The *Cysill Ar-lein Corpus* is currently password-protected and may only be accessed by researchers with the appropriate access rights due to the privacy concerns and lack of a robust Welsh-language anonymization, as reported above. The current interface was therefore initially designed to cater primarily for the internal research interests of the project team, which are mainly related to terminology, lexicography and linguistic register.

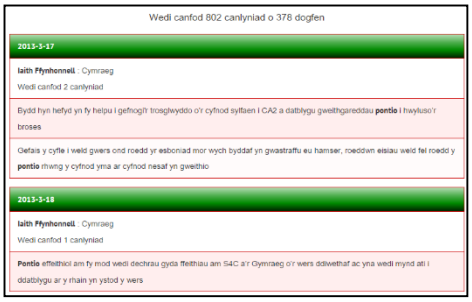

Figure 4: Cysill Ar-lein Corpus Search Results

These interests are also reflected in the publically available *Welsh National Corpora Portal* (2014) which brings together a collection of monolingual and parallel Welsh/English corpora within a simple, searchable, interface that can search for all word forms derived from a single unmutated and/or unconjugated search word. This lemmatizer, as well as a Welsh-language part of speech tagger, is also available from the *Welsh National Language Technologies Portal* (2015) as a cloud based API service. These tools were initially developed for the original *Cysill* spelling and grammar checker, and were used to analyse *Cronfa Electroneg o'r Gymraeg* (CEG), the first major electronic corpus of Welsh (Ellis et al, 2001).

Although these range of tools may be considered basic for a major language, they are still a scarce resource for many minority languages, and are invaluable in the analysis of a moderately inflected language such as Welsh. Of course, language independent corpus tools such as concordances may also be used to analyse the *Cysill Ar-lein* corpus, and *MonoConc Pro 2.2* (Barlow, 2003) was successfully used in Wooldridge's (2011) work on the corpus.

One of the features that was developed expressly for the *Cysill Ar-lein Corpus* was easy access to a list of the most common n-grams, 3-grams, 2-grams and 1-grams identified in the corpus. These lists, when filtered by the '?' (unrecognised) part of speech tag provide a list of the most frequently occurring tokens found in the corpus that were not recognised by the part of speech tagger. Whilst some neologisms have been identified in this manner, these unrecognised forms mainly correspond to informal spoken dialectal forms and reflect the fact that *Cysill Ar-lein* is

often used to check the spelling and grammar of texts that are more informal than for which it was designed. These informal, colloquial forms may not appear in standardized dictionaries in the foreseeable future, but they may, in conjunction with the n-gram statistics, help in the task of generating new language models for speech recognition in another project currently underway where this approach is now being trialled.

6. Conclusions

As a large corpus of original contemporary written Welsh the *Cysill Ar-lein Corpus* has the potential to become invaluable to research areas from neology and named entity extraction to sociolinguistics. In the minority language context, it represents a novel method of corpus collection through the provision of a valued service (a proofreading tool in this case) in exchange for the language users' data. By ensuring that data submissions directly benefit the language users, the effort required to collect large quantities of language data has been significantly reduced. Whilst such an approach does require possessing a valued service in the first place, the success of *Cysill Ar-lein* as a means of data collection might influence the prioritisation of language resources in other less resourced languages.

In terms of resulting research, the *Cysill Ar-lein Corpus* has already been used internally for lexical research into the identification of neologisms and dialectal forms of the verb, as well as for building language models to aid machine translation and speech recognition. As one of the largest, if not the largest, general language corpora available to date for Welsh, it is also a valuable source of frequency counts, not only of words, but of Welsh morphological features such as initial consonant mutation.

In both its corrected and uncorrected form the corpus gives an insight into the production of written Welsh that has not for the most part gone through the editorial process undergone by the mostly published texts found in previous Welsh-language corpora. As a result, one anticipated use of the corpus is to analyse patterns of linguistic errors in the uncorrected corpus, using the results to improve the teaching of language to learners, schoolchildren and less proficient writers of Welsh. The software developers also intend using the corrected corpus to identify any common linguistic errors that the proofreading software currently cannot correct. There are also interesting questions regarding the acceptability of certain suggested corrections that might be perceived as archaic or otherwise unwelcome by some users. By analysing user's acceptance or rejection of the suggestion, such instances could be identified and a clearer picture of the current use of the language may be established.

External researchers may discover other novel uses for the corpus, in both corrected and uncorrected forms. However, the need to develop a robust method for Welsh-language text anonymization prior to more widespread distribution is a significant hurdle that must be overcome. It is clear, despite the notice given within the Cysill service's terms and conditions, that users do not always understand the implications of allowing confidential details to be stored in such systems. In this case, we are working towards a resolution for the anonymization problem in the hope the corpus, or at least a subset of it, can be published publically.

We also hope to make greater use of the corpus for linguistic research, especially the raw uncorrected corpus, which we believe to be the only such corpus in existence for the Welsh language to date.

7. Bibliographical References

2001 Census, Key Statistics for Wales, March 2011. Office for National Statistics. http://www.ons.gov.uk/ons/rel/census/2011-census/key-statistics-for-unitary-authorities-in-wales/stb-2011-census-key-statistics-for-wales.html. Accessed 17/09/2015.

Barlow, Michael. (2003) MonoConc Pro 2.2. Houston: Aethlestan

Cysgliad (2004). Bangor University: Bangor.

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

Example Corpus of Registers (2014). Bangor University: Bangor. http://corpws.cymru/registers/?lang=en. Accessed 17/09/2015.

Ellis, N. C., O'Dochartaigh, C., Hicks, W., Morgan, M., & Laporte, N. (2001). Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh. http://www.bangor.ac.uk/canolfanbedwyr/ceg.php.en. Accessed 08/05/2014

Hicks, W.J. (2004). "Welsh Proofing Tools: Making a Little NLP go a Long Way." Proceeding of the1st Workshop on International Proofing Tools and Language Technologies. Greece: Univeristy of Patras.

Prys, Delyth (2008). "The Development and Acceptance of Electronic Resources for Welsh" Proceedings of the Second Colloquium on Lesser Used languages and Computer Linguistics. Bozen-Bolzano. Eurac Research.

Welsh National Corpora Portal (2014). http://corpws.cymru. Accessed 17/09/2015.

Welsh National Language Technologies Portal (2015). http://techiaith.cymru/api/cysill-ar-lein/?lang=en. Accessed 17/09/2015.

Wooldridge, Dawn. (2011). Gwella Cysill at Ddefnydd Cyfieithwyr: adnabod ymyrraeth gan yr iaith Saesneg mewn testunau Cymraeg. (Improving Cysill for Use by Translators; recognizing interference by English in Welsh texts). MRes Thesis, Bangor University. Published online at https://www.cyfieithwyrcymru.org.uk/files/THESIS_DAWN_WOOLDRIDGE1.pdf. Accessed 09/03/16.