Cysill Ar-lein Corpus

A Corpus of written contemporary Welsh compiled from an on-line spelling and grammar

Delyth Prys, Gruffudd Prys & Dewi Bryn Jones
d.prys@bangor.ac.uk, g.prys@bangor.ac.uk, d.b.jones@bangor.ac.uk

Introduction

- The *Cysill Ar-lein Corpus* is a 31 million word (and growing) corpus of written corpus of contemporary Welsh collected from user submissions to an online grammar and spellchecking tool. It is held at Bangor University, Wales.

Cysill Ar-lein: online grammar and spellchecking

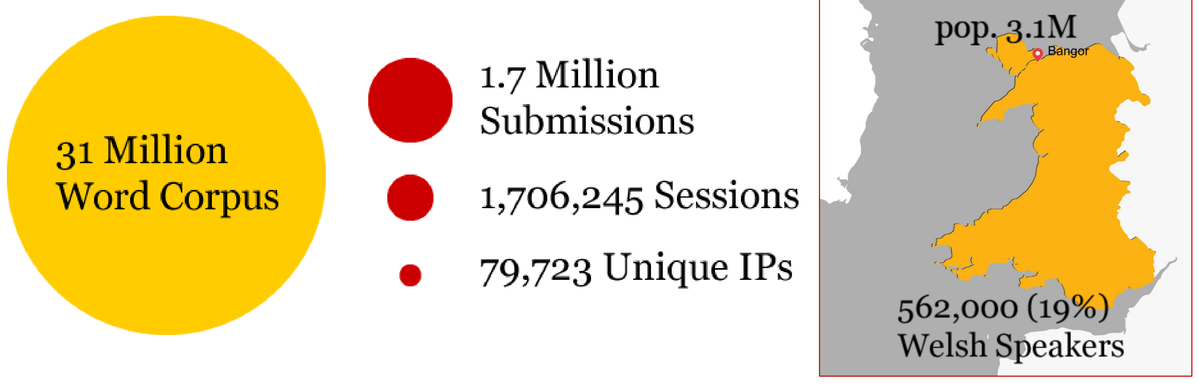
- The tool, *Cysill Ar-lein,* is a free online service based on the commercially successful *Cysgliad* software for Microsoft Windows. Limitations compared to the commercial version are:
- User submissions are limited to **3000 characters**.
- Users must accept terms of use that explicitly allow any submitted texts to be used for research purposes

Cysill Ar-lein User Interface

- Between 11/03/13 and 09/03/15:
  - On average, over 600 daily users of the website, generating approximately 1,100 daily page views.
  - Average time spent on the website was 7.13 minutes, indicating considerable time spent inputting texts, looking at suggested corrections, and either accepting or rejecting them.
  - 79,723 unique IPs, giving a total number of 1,706,245 sessions where text was input for automatic proofreading by the system.
  - Considering a total population of 562,000 Welsh speakers (2011 Census, Key Statistics for Wales, 2011) these figures are surprisingly high.

Corpus description

- Contains over 31 million words from a total of 1.7 million submissions.
- The corpus consists mainly of unedited texts chiefly produced by non-professional writers, and, as a separate corpus, the edited outcome of the proofing process.



- The corpus is a monitor corpus: it grows as users check the spelling and grammar of their written documents using the service.
- A copy of the user's text is recorded at each point during the spelling and grammar checking process, so that all spelling or grammar edits prompted by the program are recorded.
- Submitted texts include a wide range of subjects including Correspondence, Creative Writing, Academic Work, Publicity and Other.
- Author details, e.g. age, gender, language proficiency not collected due to privacy concerns, but texts are usually of sufficient length to identify user's language level and the subject matter of the text.
- Despite warning in terms and conditions, confidential information found in corpus, therefore access is currently restricted to researchers at Bangor University.
- However, manually anonymized subcorpora such as the Example Corpus of Language Registers can be distributed publicly.

Corpus tools

- Designed to cater for the research interests of the project team in terminology, lexicography and linguistic register.
- Include simple, searchable, interface based on *Welsh National Corpora Portal* (2014) with lemmatization functionality to show all word forms derived from a single unmutated and unconjugated word.
- Features the ability to generates list of the most common n-grams, 3-grams, 2-grams and 1-grams found in the corpus.
- Export functionality allows use with NLP toolkits such as NLTK
- Text can be automatically POS tagged using the cloud based Welsh POS Tagging API service from the *Welsh National Language Technologies Portal* (2015).

Conclusions

- In the minority language context, the Cysill Ar-lein Corpus represents a novel method of corpus collection, where language users bring their texts to the researchers rather than the researchers having to source texts from the users.
- The result is one of the largest general language corpora to date for Welsh.
- The submission of unproofed, unedited texts gives insight into the production of written Welsh by less proficient users.
- The collection and comparison of uncorrected and corrected corpus allows the identification of any common linguistic errors that the proofreading software currently cannot correct.
- The need to anonymize data before making it more widely available is an issue that needs resolving without compromising users' trust in the spelling and grammar checker.
- The *Cysill Ar-lein Corpus* has provided researchers with a valuable large-scale corpus at very little additional cost or effort.

References

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

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.

PRIFYSGOL BANGOR UNIVERSITY