CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes– the National Corpus of Contemporary Welsh

Project overview:

The CorCenCC corpus contains over 11 million words (circa 14.4m tokens) from written, spoken and electronic (online, digital texts) Welsh language sources, taken from a range of genres, language varieties (regional and social) and contexts. The contributors to CorCenCC are representative of the over half a million Welsh speakers in the country. The creation of CorCenCC was a community-driven project, which offered users of Welsh an opportunity to be proactive in contributing to a Welsh language resource that reflects how Welsh is currently used.

To make CorCenCC as representative of contemporary Welsh as possible, the project team designed a bespoke sampling framework. Extracts were collected from sources including for example, journals, emails, sermons, road signs, TV programmes, meetings, magazines and books. Conversations were recorded by the research team, and a specially designed crowdsourcing app (see: https://www.corcencc.org/app/) enabled Welsh speakers in the community to record and upload samples of their own language use to the corpus. The published corpus therefore contains data from Welsh speakers from all kinds of backgrounds, abilities and contexts, capturing how Welsh is truly used today across the country.

A beta version of some bilingual corpus query tools have also been created as part of the CorCenCC project (see: www.corcencc.org/explore). These include simple query, full query, frequency list, n-gram, keyword and collocation functionalities. The CorCenCC website also contains Y Tiwtiadur, a collection of data-driven teaching and learning tools designed to help supplement Welsh language learning at all different ages and levels. Y Tiwtiadur contains four distinct corpus-based exercises: Gap Filling (Cloze), Vocabulary Profiler, Word Identification and Word-in-Context. To access this tool, visit: https://www.corcencc.org/y-tiwtiadur/

The CorCenCC project was led by Dawn Knight (KnightD5@cardiff.ac.uk), at the Centre for Language and Communication Research, Cardiff University. The full project team comprised: 1 Principal Investigator (PI – Dawn Knight), 2 Co-Investigators (CIs – Steve Morris and Tess Fitzpatrick), who made up, with the PI, the CorCenCC Management Team, a total of 7 other CIs and 8 Research Assistants/Associates over the course of the project. In addition, there were 11 advisory board members, 6 consultants (from 4 countries around the world), 2 PhD students, 4 Undergraduate summer placement students, 4 professional service support staff, 4 project ambassadors and 2 project volunteers. More information can be found on the project website: www.corcencc.org

Dataset:

The CorCenCC dataset includes 14,338,149 tokens (circa 11.2-million-words). The data in CorCenCC represents a wide range of contexts, genres and topics. This data has, as far as possible, been anonymised using a combination of manual and automated techniques, and has been fully tagged in terms of part-of-speech (POS) and semantic categories. The POS and semantic tagging was carried out using CyTag and SemCyTag tools, available from CorCenCC's GitHub website: https://github.com/CorCenCC

The following files are included in this dataset:

1. **categorisation_guide**: guide to interpreting columns in CorCenCC's corpus tables/files.
2. **categorization**: links individual contribution_id's to specific taxonomy_id's (from the corpus design frame). Refer to taxonomy file for details.

3. **complete_corpus**: zipped folder containing all individual contribution files (data is fully POS and semantic tagged).
4. **contrib_links**: linking specific contributor_id's to individual contributions.
5. **contribution**: list of all contributions in the corpus (linking to specific modes).
6. **contributor**: contributor metadata for the complete corpus.
7. **corpus_data**: fully POS and semantically tagged CorCenCC corpus data.
8. **electronic**: metadata associated with individual contribution_id's (electronic mode).
9. **spoken**: metadata associated with individual contribution_id's (spoken mode).
10. **taxonomy**: metadata taxonomy guide, used as a basis for classifying contributions according to their genre, context, location, target audience, topic, who (i.e. interlocutors), and source.
11. **written**: metadata associated with individual contribution_id's (written mode).

The dataset is available from: http://doi.org/10.17035/d.2020.0119878310

License and acknowledgements:

The research on which this dataset, the accompanying software tools, and online corpus resource, are based was funded by the UK Economic and Social Research Council (ESRC) and Arts and Humanities Research Council (AHRC) as the *Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh): A community driven approach to linguistic corpus construction* project (Grant Number ES/M011348/1).