From Corpus to Community: New NLP Tools for Welsh Language Research and Learning

Dawn Knight[1], Fernando Alva-Manchego[2]

[1]School of English, Communication and Philosophy; [2]School of Computer Science and Informatics, Cardiff University, UK
[1]KnightD5@cardiff.ac.uk, [2]AlvaManchegoF@cardiff.ac.uk

Abstract

Launched in 2020, CorCenCC (*Corpws Cenedlaethol Cymraeg Cyfoes* – National Corpus of Contemporary Welsh) is the first large-scale corpus of the Welsh language to integrate spoken, written, and electronically mediated data, offering a comprehensive snapshot of contemporary Welsh use. Including contributions from over 2,000 speakers, the 11.2-million-word corpus represents the diversity of Wales's linguistic landscape. As a national resource, CorCenCC enables users to explore real world Welsh. Several tools and resources were developed through the CorCenCC project, including the CyTag POS tagger and CySemTag (adapted from Lancaster University's USAS semantic system), to enable the grammatical and semantic categorisation of the dataset. The team also built the pedagogic toolkit *Y Tiwtiadur*, to allow learners and teachers to access corpus-based examples and tasks. Additionally, *Yr Amliadur* provides curated frequency-based wordlists across modes and parts of speech, supporting linguistic analysis and vocabulary development. Since completing the corpus, the team has focused on extending its impact and reach, to ensure that the resources are maintained and sustained for future use; a challenge often faced when large-scale projects end. This poster profiles the tools and resources created from and inspired by CorCenCC and its associated tools and resources, as a means of supporting the democratisation of linguistic resources for minoritised language contexts.

Keywords: CorCenCC, national corpus, NLP, Welsh, FreeTxt, *Proffiliadur, Y Tiwtiadur, Yr Amliadur*

1. Introduction

Launched in 2020, CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – National Corpus of Contemporary Welsh) is the first large-scale corpus of the Welsh language to integrate spoken, written, and electronically mediated data, offering a comprehensive snapshot of contemporary Welsh use. Including contributions from over 2,000 speakers, the 11.2-million-word corpus represents the diversity of Wales's linguistic landscape: regions, demographics, text types, modes, and genres. As a national resource, CorCenCC enables users to explore real-world Welsh, supporting research, teaching, lexicography, and translation (Knight et al, 2020a; Knight et al., 2020b). A screenshot of CorCenCC's query tools is provided in Figure 1.

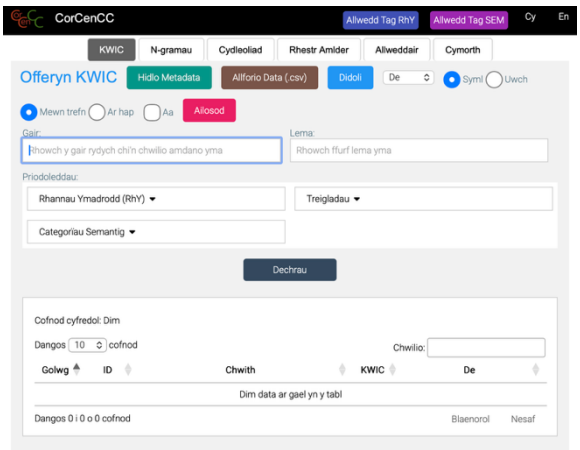

Figure 1: CorCenCC's query tools.

Several tools and resources were developed through the CorCenCC project, including the CyTag POS tagger (Neale et al., 2018) and CySemTag (adapted from Lancaster University's USAS semantic system – Piao et al., 2018), to enable the grammatical and semantic categorisation of the dataset. The team also built the pedagogic toolkit *Y Tiwtiadur*, to allow learners and teachers to access corpus-based examples and tasks. Figure 2 depicts the Word Identifier (*Abnadbod Geiriau*) functionality in the *Y Tiwtiadur* toolkit.

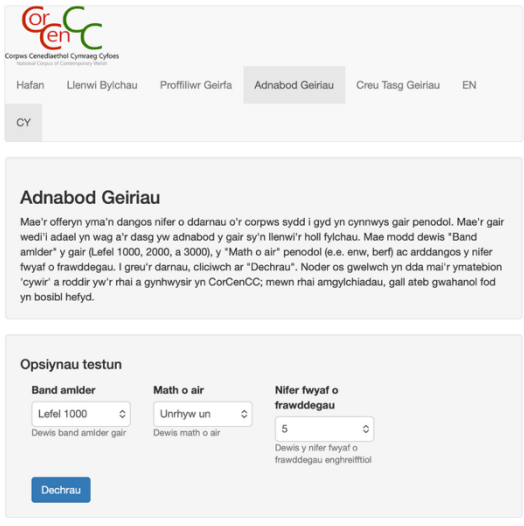

Figure 2: *Y Tiwtiadur*'s Word Identifier tool.

Additionally, *Yr Amliadur* provides curated frequency-based wordlists across modes and parts of speech, supporting linguistic analysis and vocabulary development (Knight et al. 2023).

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 101–103. Palma, Mallorca. May 11, 2026.

2. Extending CorCenCC's reach

Since completing the corpus, the team has focused on extending its impact and reach, to ensure that the resources are maintained and sustained for future use; a challenge often faced when large-scale projects end. This poster profiles the tools and resources created from and inspired by CorCenCC and its associated tools and resources, as a means of supporting the democratisation of linguistic resources for minoritised language contexts.

First, the poster profiles the development of *Geirfan*, a pedagogic wordlist created through a partnership with the National Centre for Learning Welsh, the Welsh Joint Education Committee, and language experts. Building on *Yr Amliadur*, the team developed frequency-driven vocabulary lists tailored to A1-level adult learners and created a prototype online dictionary. Since 2022, around 1,600 candidates have taken WJEC assessments based on *Geirfan* resources annually, marking the first time Welsh for Adults curricula have drawn directly on corpus-derived frequency data.

The poster also profiles the Welsh Government-funded ACC Welsh Automatic Text Summarisation tool (El-Haj et al., 2022; Ezeani et al., 2022), which allows users to generate concise summaries of long Welsh texts, supporting teaching and public access to information. Combining ACC, the CorCenCC dataset and its taggers, FreeTxt (Knight et al., 2024), is another resource developed by members of the team, as seen in Figure 3. FreeTxt enables Welsh and English qualitative data analysis using corpus-based NLP methods in an accessible interface co-designed with major Welsh cultural and educational organisations.

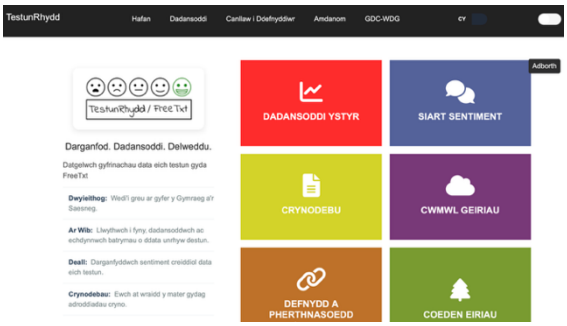

Figure 3: FreeTxt

Furthermore, *Proffiliadur,* our Python-based readability toolkit (Gutiérrez-Rolón et al., 2026), provides the first dedicated text-profiling resource for Welsh, offering reproducible, linguistically grounded measures of readability in a low-resource language.

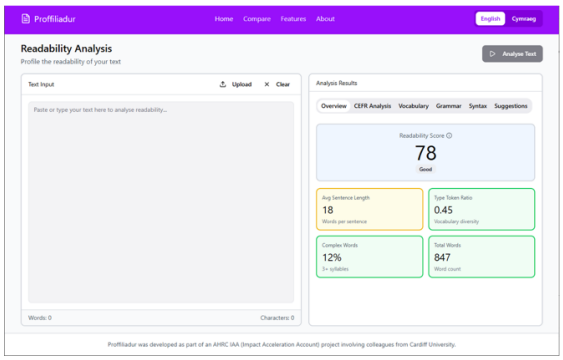

Figure 4: *Proffiliadur*

Proffiliadur computes 141 surface, lexical, morphological, and syntactic indices, designed to capture linguistic variation while incorporating Welsh-specific processing that enables accurate morphological analysis and handles phenomena such as initial consonant mutation. Proffiliadur enables assessment of text accessibility, supporting applications in education, healthcare, and public communication.

Finally, Gutierrez-Rolón and Alva-Manchego (2026) developed a mutation trigger identifier using, among other resources, the CyTag POS tagger and CorCenCC's query tool. The latter was instrumental in identifying examples of various types of Welsh mutations in real-world usage.

3. Sustainability and next steps

Through developing of these tools, the team has worked to empower end-users to direct and lead their own analyses of both small-scale and more extensive qualitative datasets to maximise the reach and potential impact. The approaches used to construct the resources serve as a template for those seeking to develop corpora and democratise language technology use in other minoritised and major language contexts around the world (e.g. Nguyen et al., 2026).

All resources are freely accessible via our Welsh Government funded DigiGrid platform (see Figure 4), which brings together a suite of tools to support Welsh language exploration, learning, and analysis.

Figure 5: The DigiGrid platform.

4. Acknowledgments

The work reported on in this poster is broadly based on the ESRC (Economic and Social Research Council) and AHRC funded Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh): A community driven approach to linguistic corpus construction project (ES/M011348/1). The FreeTxt project was funded by AHRC (Arts and Humanities Research Council) follow-on funding for impact and engagement (AH/W004844/1). The development of the *Geirfan* wordlists and *Proffiliadur* toolkit were funded by Cardiff University's AHRC IAA account. Finally, the ACC Welsh Automatic Text Summarisation tool and DigiGrid platform were funded by Welsh Government's Welsh Language Technology grants.

5. Bibliographical References

El-Haj, M., Ezeani, I., Morris, J. and Knight, D. (2022). Creation of an evaluation corpus and baseline evaluation scores for Welsh text summarisation. *Proceedings of the Celtic Language Technology Workshop, LREC (Language Resources Evaluation) 2022 Conference*, June 2022, Marseille, France.

Ezeani, I., El-Haj, M., Morris, J., & Knight, D. (2022). Introducing the Welsh text summarisation dataset and baseline systems. *Proceedings of the LREC (Language Resources Evaluation) 2022 Conference*, June 2022, Marseille, France.

Gutiérrez-Rolón, N., Davies, J., Williams, T., Knight, D. & Alva-Manchego, F. (2026). Proffiliadur: Welsh Language Text Profiling Toolkit. *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (LREC), May 2026, Palma de Mallorca, Spain.

Gutiérrez-Rolón, N. & Alva-Manchego, F. (2026). Unsupervised Labelling of Mutation Triggers in Welsh. *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (LREC), May 2026, Palma de Mallorca, Spain.

Knight, D., Loizides, F., Neale, S. Anthony, L., & Spasić, I. (2020a). Developing computational infrastructure for the CorCenCC corpus – the National Corpus of Contemporary Welsh. *Language Resources and Evaluation, 55*(1)*,* 1-28.

Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., Thomas, E-M., Lovell, A., Morris, J., Evas, J., Stonelake, M., Arman, L., Davies, J., Ezeani, I., Neale, S., Needs, J., Piao, S., Rees, M., Watkins, G., Williams, L., Muralidaran, V., Tovey-Walsh, B., Anthony, L., Cobb, T., Deuchar, M., Donnelly, K., McCarthy, M. and Scannell, K. (2020b). CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh. Cardiff University. http://doi.org/10.17035/d.2020.0119878310

Knight, D., Fitzpatrick, T., Morris, S., Tovey-Walsh, B., Prosser, H., & Davies, E. (2023). Corpus to curriculum: Developing word lists for adult learners of Welsh. *Applied Corpus Linguistics, 3*(2), article number: 100052.

Knight, D., Khallaf, N., El-Haj, M., Ezeani, I., & Morris, S. (2024). FreeTxt: a corpus-based bilingual free-text survey and questionnaire data analysis toolkit. *Applied Corpus Linguistics, 4*(3), article number: 100103.

Neale, S., Donnelly, K., Watkins, G., & Knight, D. (2018). Leveraging lexical resources and constraint grammar for rule-based part-of-speech tagging in Welsh. *Proceedings of the LREC (Language Resources Evaluation) 2018 Conference*, May 2018, Miyazaki, Japan.

Nguyen, H. H., El-Haj, M., Rayson, P., & Knight, D. (2026). FreeTxt-Vi: A Benchmarked Vietnamese-English Toolkit for Segmentation, Sentiment, and Summarisation. Proceedings of *Learning Resources Evaluation Conference 2026* (LREC), May 2026, Palma de Mallorca, Spain.

Piao, S., Rayson, P., Knight, D., & Watkins, G. (2018). Towards a Welsh semantic annotation system. *Proceedings of the LREC (Language Resources Evaluation) 2018 Conference*, May 2018, Miyazaki, Japan.