The National Corpus of Contemporary Welsh

Project Report, October 2020

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

1. Introduction

1.1. Purpose of this report

This report provides an overview of the CorCenCC project and the online corpus resource that was developed as a result of work on the project. The report lays out the theoretical underpinnings of the research, demonstrating how the project has built on and extended this theory. We also raise and discuss some of the key operational questions that arose during the course of the project, outlining the ways in which they were answered, the impact of these decisions on the resource that has been produced and the longer-term contribution they will make to practices in corpus-building. Finally, we discuss some of the applications and the utility of the work, outlining the impact that CorCenCC is set to have on a range of different individuals and user groups.

1.2. Licence

The CorCenCC corpus and associated software tools are licensed under Creative Commons CC-BY-SA v4 and thus are freely available for use by professional communities and individuals with an interest in language. Bespoke applications and instructions are provided for each tool (for links to all tools, refer to section 10 of this report). When reporting information derived by using the CorCenCC corpus data and/or tools, CorCenCC should be appropriately acknowledged (see 1.3).

- To access the corpus visit: www.corcencc.org/explore
- To access the GitHub site: https://github.com/CorCenCC
    - GitHub is a cloud-based service that enables developers to store, share and manage their code and datasets.

1.3. Referencing CorCenCC

Appropriate credit needs to be given when using the CorCenCC corpus data and/or tools. To reference the CorCenCC corpus and the present project report, please use the following:

- **CorCenCC corpus**: 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. (2020). CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh. Cardiff University. http://doi.org/10.17035/d.2020.0119878310

- **Report:** Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I. and Thomas, E. M. (2020). The National Corpus of Contemporary Welsh: Project Report | Y Corpws Cenedlaethol Cymraeg Cyfoes: Adroddiad y Prosiect. *arXiv:2010.05542*, October 2020.

Other project publications can be found in section 10 of this report and on the 'Outputs' tab of the CorCenCC website: www.corcencc.org/outputs

1.4. Acknowledgements

The research on which this report, and the accompanying 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). Information about project team members can be found at www.corcencc.org/contacts. Without their input, expertise, enthusiasm and collegiality, the CorCenCC project would not have been possible.

We would also like to acknowledge Cardiff University and Swansea University for their contribution of PhD scholarships, enabling us to include postgraduate researchers in the project team. We particularly extend personal thanks to those colleagues at all our respective universities who have generously given their time and support to us at critical stages of the project.

The fulfilment of the CorCenCC project is also thanks to our project stakeholders (see 2.2), and especially those in the project advisory group: not only have they been generous in facilitating the collection of, or directly contributing, data, they have also been encouraging in their support for the aims of the project and their engagement with the planning process, as well as their commitment to the sustainability and continuation of CorCenCC.

2. Vision and objectives

2.1. Project overview

CorCenCC is an inter-disciplinary and multi-institutional project that has created a large-scale, open-source corpus of contemporary Welsh. A corpus, in this context, is a collection of examples of spoken, written and/or e-language examples from real life contexts, that allows users to identify and explore language as it is actually used, rather than relying on intuition or prescriptive accounts of how it 'should' be used. Corpora let us investigate how we use language across different genres and communicative mediums (i.e. spoken, written or digital), and how it varies according to the speaker/writer and the communicative purpose. This evidence-based approach is used by academic researchers, lexicographers, teachers, language learners, assessors, resource developers, policy makers, publishers, translators and others, and is essential to the development of technologies such as predictive text production, word processing tools, machine translation, speech recognition and web search tools.

Prior to the construction of CorCenCC, there were a number of Welsh language corpora in existence, including the 460k word spoken Siarad corpus (Deuchar et al., 2018),

the 24m word e-language-based Crúbadán Welsh Corpus (Scannell, 2007) and the crowdsourced 40-hour Paldaruo Speech Corpus comprised of read-aloud texts (Cooper et al., 2019). Consideration was given as to whether these could be integrated or aligned with the objectives of CorCenCC. Deuchar and Scannell were engaged as consultants for the project, while Canolfan Bedwyr (Bangor) was represented on the project advisory group. However, given that the previous corpora were compiled to realise different, distinct and bespoke aims and visions, it was considered necessary to create a new, complete dataset.

CorCenCC is the first corpus of the Welsh language that covers all three aspects of contemporary Welsh: spoken, written and electronically mediated (e-language). It offers a snapshot of the Welsh language across a range of contexts of use, e.g. private conversations, group socialising, business and other work situations, in education, in the various published media, and in public spaces. It includes examples of news headlines, personal and professional emails and correspondence, academic writing, formal and informal speech, blog posts and text messaging (the specific composition of the corpus is discussed in section 3.3). Language data was sampled from a range of different speakers and users of Welsh, from all regions of Wales, of all ages and genders, with a wide range of occupations, and with a variety of linguistic backgrounds (e.g. how they came to speak Welsh), to reflect the diversity of text types and of Welsh speakers found in contemporary Wales. In this way, the CorCenCC corpus provides the means for empowering users of Welsh to better understand and observe the language across diverse settings, and creates a solid evidence base for the teaching of contemporary Welsh to those who aspire to use it. Over time, the corpus has the potential to make a significant contribution to the transformation of Welsh as the language of public, commercial, education and governmental discourse.

To that end, CorCenCC is designed to enable, for example, community users to investigate dialect variation or idiosyncrasies of their own language use; professional users to profile texts for readability or develop digital language tools; Welsh language learners to draw on real life models of Welsh; and researchers to investigate patterns of language use and change. The corpus is also anticipated to reveal new insights into the vocabulary and language patterns of Welsh and to serve as a major resource for teaching the Welsh language to both those who have it as their first language and new speakers of it. This multifaceted impact potential has been made possible by CorCenCC's significant contribution at the methodological level, in extending the scope, relevance and design infrastructure of language corpora. Specifically, the project has involved the development of important new tools and processes, including a unique user-driven corpus design in which language data was collected and validated through crowdsourcing, and an in-built pedagogic toolkit (Y Tiwtiadur) developed in consultation with representatives of all anticipated academic and community user groups (for a detailed discussion of CorCenCC's user-driven design, see Knight et al., 2021a, Knight et al., 2021b – details in section 10.2. below).

2.2. Project team

The CorCenCC project involved 4 academic institutions (Cardiff, Swansea, Lancaster and Bangor Universities), 1 Principal Investigator (PI – Dawn Knight), 2 Co-Investigators (CIs – Tess Fitzpatrick and Steve Morris) who made up, with the PI, the CorCenCC Management Team (CMT), a total of 7 other CIs (Irena Spasić, Paul Rayson, Enlli Môn Thomas, Alex

Lovell, Jonathan Morris, Jeremy Evas, Mark Stonelake), 10 Research Assistants/Associates (RAs), and 180+ transcribers working over the course of the project.

In addition, there were 6 consultants, 2 PhD students, 4 undergraduate summer placement students, 4 professional service support staff and 2 project volunteers. The project also benefitted from contributions and support from representatives of a range of stakeholders including the Welsh Government, National Assembly for Wales, BBC, S4C, WJEC, Welsh for Adults, Gwasg y Lolfa, SaySomethinginWelsh and University of Wales Dictionary of the Welsh Language, via a Project Advisory Group (PAG). Nia Parry (TV presenter, producer and researcher; Welsh tutor, *Welsh in a week* (S4C)), Nigel Owens (international rugby referee; TV presenter), Cerys Matthews (Musician author; radio and TV presenter) and Damian Walford Davies (poet; professor of English and Welsh Literature; former Chair of Literature Wales) are the official ambassadors of the CorCenCC project. A full list of all individuals involved in the project can be found at www.corcencc.org/contacts - and many are referred to throughout this report, as relevant.

The project was facilitated by a strong cross-institution team, which supported staff recruitment, financial management, information technology (including equipment, software, project server maintenance and websites), media and communications outreach (including press release coordination, and radio and TV appearances), legal guidance on forms, contracts and licences, and Welsh language translators and interpreters (providing written translations of reports, key project documents and other outputs, and live simultaneous interpreting during the Whole Project Team (WPT) meetings and public dissemination events). Over 210 (bi)weekly reports were written across the course of the project, detailing work completed, issues and key risks, work to be undertaken, new ideas, thoughts and opportunities; ten Whole Project Meetings were held, and well over 100 additional meetings took place. Seven internal mailing lists were created to enable communication between team members located at different sites, along with a central, whole project Gantt chart, to collaboratively document and track deliverables and key project milestones. To maintain communication with the general public and other stakeholders, 24 editions of a project newsletter were published, circulated to individuals, and uploaded to the main website. The newsletters updated readers on data collection, reported on presentations and keynote presentations delivered (of which there were 54 in total), and introduced them to individual members of the team via our regular 'meet the team' slot. The function of this was to sustain interest and a sense of investment in the project (in line with the user-driven design). The project websites (www.corcencc.org | www.corcencc.cymru), Facebook and Twitter feeds also facilitated public engagement, amassing over 140,000 website hits and gaining 1029 followers on Twitter, 374 on Facebook (to August 2020). Although mentioned last here, the most important members of the extended CorCenCC team are the 2000+ individual contributors to the corpus.

2.3. Work Packages (WPs)

Work on the CorCenCC project was distributed across six coordinated work packages (WPs), each with specific tasks, aims and objectives. Led by Knight, WP0 attended to the on-going design, scoping and training activities, and involved all members of the project team. The other WPs were:

- WP1: Collect, transcribe and anonymise the data
- WP2: Develop the part-of-speech tag-set/tagger
- WP3: Develop a semantic tagger for Welsh and semantically tag all data
- WP4: Scope, design and construct Y Tiwtiadur
- WP5: Construct the infrastructure to host CorCenCC and build the corpus

While work was distributed across these work-packages, colleagues had a mutual understanding of the shared vision for the project and worked collaboratively to achieve it, with a considerable measure of interdependence between WPs that required discussion and coordination. For example, WP3 built on the research undertaken in WP1 for corpus collection, and employed WP2's part of speech (POS) tagger as a first step in the semantic analysis of the Welsh language data. WP3's output then fed into WP4 for the online pedagogic toolkit (Y Tiwtiadur), which used the multiple levels of corpus annotation to improve the engagement with and affordances of the toolkit for teachers and learners. Additionally, WP3's semantically tagged corpus fed directly into the corpus infrastructure developed in WP5. In the following sections we provide a detailed description of the WPs, outline their main aims and objectives, and reflect on their key achievements, contributions and potential applications. These descriptions were written by the respective WP leads.

3. Work Package 1: Collect, transcribe and anonymise the data

3.1. WP1: Description

The main work of WP1 concerned the sourcing, collecting and processing of the data to be included in CorCenCC. Core elements of this procedure were i) the creation of the project's sampling frame; ii) establishing transcription conventions; iii) ensuring a uniform approach to ethical compliance in the collection of the data. WP1 was co-led by Morris and Knight, who were joined by a team of Welsh speaking researchers, including CIs Evas, J. Morris, and Lovell, and RAs Needs, Rees, Arman, Watkins and Williams (at differing points throughout the project). Deuchar and McCarthy, leaders in the field of corpus linguistics, provided on-going consultative advice for this phase throughout the project and additional support was provided by a number of project volunteers and interns.

3.2. WP1: Objectives

The aims and objectives for WP1 were

- to design a sampling frame for the corpus
- to source and collect appropriate data
- to design and apply transcription protocols to spoken data

Sampling frame

The project proposal's Case for Support included an outline guide to our objectives regarding language modes (spoken, written, e-language), genres and topics, with approximately how many words of data would be collected under each heading. One of the first tasks in WP1 was to further refine and develop this outline guide. A sampling frame was created to

underpin the data collection for the project, to ensure that we captured a range of different speakers across different discourse contexts and geographical locations. The sampling frame was designed to reflect current demographics of Welsh speakers using up-to-data census information (ONS, 2011). An innovative aspect of the CorCenCC sampling frame is the detailed consideration of domains in which the Welsh language is used. In a context where the great majority of Welsh speakers are bilingual and there is an uneven geographical spread in terms of density of speakers, age of speakers and language domains, the sampling frame needed to reflect the contemporary sociolinguistic situation of the language as accurately as possible.

Sourcing the data

The targets for, and sources of, the spoken, written and e-language data to be collected for CorCenCC were driven by our sampling frame and were shaped by initial investigation of where contemporary Welsh language is spoken and where written and e-language material is concentrated. In a bilingual context, certain domains may be under-represented (e.g. national daily newspapers). It was therefore necessary to ensure that the data was a true reflection of what is available and accessible to users of the language, rather than replicating frameworks designed for creating corpora in languages where the majority of the speakers are monolinguals.

Transcription

There were two preparatory steps: (i) the creation of transcription conventions for Welsh and (ii) the recruitment of transcribers (see also 3.3 below). There were particular challenges in step (i) for a number of reasons:

- In the written language, authors often denote variation in spoken varieties, so that the same basic linguistic meaning is written in many different ways. An example would be the first person singular present tense of 'to be'. In English, this might be realised as '*I am*' or '*I'm*'. Formal written Welsh would give '*Yr wyf (i)*' or '*Rwyf i*'. However, spoken Welsh produces some of the following possibilities: '*Rydw i / Dw i / Rwy / Wy / Fi*'. Writers would use these forms to represent speakers from different areas and they can be observed in literature and in other written media. The transcription conventions therefore needed to be able to reflect this reality while indexing it to the same meaning, so that searches could find the different realisations.
- Given that CorCenCC includes electronic language as well as written and spoken data, and that the conventions for representing language in the electronic context are not fully established in any language, it was necessary to be similarly accommodating in these cases.
- With regard to transcribing spoken material, the general principle adopted was to align what was heard with the closest written realisation from a set that began with the existing range of written forms but was augmented as necessary to ensure all spoken forms were appropriately captured. This principle was refined and developed through several iterations of the CorCenCC transcription conventions.

- Having adopted this general transcription principle, there was a need to ensure consistency across the CorCenCC transcriber pool. Any tendency towards prescriptivism or defaulting to formal written Welsh had to be robustly resisted.

Transcribers were recruited through campaigns particularly targeted at members of the translation profession (through their association's newsletter), students at university and those who had been involved in transcribing for other projects. Every transcriber had to pass a preliminary test piece (adhering to the CorCenCC transcription conventions) and quality was maintained through randomised checking of 25 per cent of all transcribed work with remedial corrections made where issues were identified.

3.3. WP1: Achievements

Sampling frame

Table 1 presents the initial sampling frame that was proposed in the project proposal's Case for Support. These initial distributions were based on the desirability of raising the representation of extemporised language (spoken and e-language) relative to prepared language (written).

Table 1. Initial sampling frame for the CorCenCC corpus (taken from the Case for Support).

Type	Example sources (approximate)	Words	Total
Spoken	Welsh learner discourse	600,000	**4m**
	Conversations with friends; with family; televised interviews and TV chat shows (BBC); workplace Welsh	400,000 each	
	BBC radio shows; service encounters	400,000 each	
	Phone calls; primary, secondary, tertiary and adult classroom interaction; political speeches; formal and informal interaction at the National Eisteddfod	250,000 each	
Written	Welsh learner writing	600,000	**4m**
	Books; papurau bro (i.e. community newspapers); political documents; stories	400,000 each	
	Letters and diaries; academic essays; academic textbooks; magazines; adverts; flyers/information leaflets; formal letters	290,000 each	
	Signs	60,000	
E-Language	Discussion boards; emails; blogs	500,000 each	**2m**
	Websites; tweets	300,000	
	Text messages	200,000	
		10,000,000	

Following the commencement of the project, more detailed sampling frames were developed, for the spoken (Table 2), written (Table 3) and e-language (Table 4) components of the corpus. This iterative refinement of the sampling frame was informed by the thematic groupings and discourse categorisations of existing major corpora including BNC 1994, the Spoken BNC 2014, CANCODE and CANELC (see Aston and Burnard, 1997, McEnery et al., 2017, Carter and McCarthy, 2004 and Knight et al., 2013). The frameworks of these

existing corpora provided a useful point of departure for examining to what extent the same groupings and categorisations are usable in the minoritised language context. More fine-grained information about each of the sub-genres, and justifications for these proposed distributions, is available in Knight et al., 2021 (see section 10.2.).

Table 2. Broad revised sampling frame for the Spoken element of the CorCenCC corpus.

Contexts	% of sub-corpus	Word count
Cyhoeddus/Sefydliadol / *Public/Institutional*	10%	400,000
Cyfryngau / *Media*	15%	600,000
Trafodol / *Transactional*	10%	400,000
Proffesiynol / *Professional*	10%	400,000
Pedagogaidd / *Pedagogical*	10%	400,000
Cymdeithasu / *Socialising*	22.5%	900,000
Preifat / *Private*	22.5%	900,000
	100%	**4,000,000**

Table 3. Broad revised sampling frame for the Written element of the CorCenCC corpus.

Sources	% of sub-corpus	Word count
Llyfrau / *Books*	41.75%	1,670,000
Cylchgronau, Papurau Newydd, Cyfnodolion *Magazines, Newspapers, Journals*	19.25%	770,000
Deunydd amrywiol / *Miscellaneous material*	39%	1,560,000
	100%	**4,000,000**

Table 4. Broad revised sampling frame for the e-Language element of the CorCenCC corpus

Sources	% of sub-corpus	Word count
Blog	30%	600,000
Gwefan / *Website*	30%	600,000
Ebost / *Email*	20%	400,000
Negeseuon Testun Electronig Byr / *Short Electronic Text Messages*	20%	400,000
	100%	**2,000,000**

While the sampling frame acted as an approximate guide for data collection - an 'ideal' as it were - it is rare that a final, completed corpus, mirrors the composition of the sampling frame (for further discussions of this see Hawtin, 2018). A variety of factors influenced the final composition of the corpus, including access to specific individuals and/or data types, permissions and more practical issues concerning the amount of time it takes to process specific types of data, and the extent to which this is predictable. Once all of these factors had played out, the composition of the corpus was as seen in Table 5.

The corpus is over 11,000,000 words in size, but the composition has changed so that the spoken element contains just over 2,800,000 words. While smaller than originally planned, this sub-corpus, *in itself*, is the biggest naturally occurring spoken Welsh corpus in existence. For a detailed breakdown of the corpus (including specific contexts, genres and topics, and demographic metadata categories and their definitions) see Knight et al., 2021 (see section 10.2.).

Note that the online corpus query tools provide an overall count of 14,338,149 **tokens** in the corpus, and makes calculations based on this value. Tokens are the smallest unit contained within a corpus, which includes words (i.e. items starting with a letter of the alphabet) and nonwords (i.e. items starting with a character that is not a letter of the alphabet). Corpora, therefore, always contain more tokens than words. The values discussed in this chapter are based on words only as this arguably provides a more accurate account of the **units of meaning** contained within the corpus.

Table 5. Approximate final composition of CorCenCC.

SPOKEN			
spoken_context	**No. of texts**	**No. of words**	**Total**
broadcast	564	750,078	1,331 texts 2,860,095 words
educational	136	296,709	
private	92	240,719	
professional	80	477,983	
public or institutional	137	433,361	
social	131	456,487	
transactional	191	204,758	
WRITTEN			
written_genre	**No. of texts**	**No. of words**	**Total**
academic_journal	10	304,447	704 texts 3,934,082 words
book	137	1,928,582	
essays_coursework_and_exams	31	26,047	
leaflet_document_announcement	338	800,030	
letter	53	12,873	
magazine	80	329,203	
miscellaneous	5	8,251	
newsletter	33	78,803	
papur_bro	13	117,334	
thesis	4	328,512	
E-LANGUAGE			
elanguage_genre	**No. of texts**	**No. of words**	**Total**
blog	48	2,345,909	9,397 texts 4,402,003 words
email	781	141,554	
SMS	8,487	93,541	
website	81	1,820,999	
	11,432	**11,196,180**	

Sourcing data

When recruiting contributors of spoken data, the aim was to ensure representation from all areas of Wales. Spoken data was sourced via two main approaches: (i) recruitment of participants to be recorded and (ii) recruitment of participants to contribute spoken data via the CorCenCC app (see also 3.3.). The scope of (i) included not only research assistants going into the field to record speakers but also participants recording themselves in various interactions. This was facilitated through a network of local 'champions' (active language animateurs in targeted areas) or the *Mentrau Iaith* (each local authority in Wales has an

associated *Menter Iaith*, i.e. community-based organisation dedicated to raising the profile of the Welsh language local language initiatives). Recruitment for (ii) was achieved by publicising the app (for example through social media, television appearances and publicity materials) to endeavour to reach a different cohort of participants who would be recording individually and in more private domains. Large Welsh language events such as the National Eisteddfod and Tafwyl provided opportunities for the team to reach a large cross-section of participants as well as raise general awareness of the project.

Recruiting participants to contribute using the CorCenCC phone app proved challenging. The app was made available on IoS, Android and via a web-interface (to accommodate those who did not have access to a mobile phone), and campaigns in the media e.g. appearance on television programmes such as S4C's *Prynhawn Da*, on both Welsh and English medium radio and through local engagement events generated much initial enthusiasm. However, this did not translate into a great deal of uptake of the app. Feedback from partners at the *Mentrau Iaith* suggested that people might be disproportionately concerned about being identifiable (despite frequent assurances of anonymisation), due to the relatively small size of the language community.

Promotional material (aimed at encouraging participation but also as an effective way of raising awareness of the project) included pens, coasters, leaflets and postcard size information sheets. An 'unofficial' mascot - based on a cat called Cor-pws - was designed to facilitate the participation of those under 18 and proved popular with contributors of all ages. Facebook and Twitter accounts for CorCenCC were set up in the first months of the project to further enhance the recruitment and participation of contributors.

In terms of written data, the good relationship forged at the beginning of the project with Welsh language publishers such as *Gwasg y Lolfa* led to the incorporation into the corpus of many up to date novels and books. A unique source of written data in the Welsh language is the locally based *Papurau Bro* (i.e. local community Welsh-language newspapers). It was decided to work with our local *Mentrau Iaith* contacts to collect these. Fairly rapid data capture, for example, sampling from the Welsh language academic journal *Gwerddon* through the *Coleg Cymraeg Cenedlaethol* and adult L2 pedagogical resources / examination papers through the *Welsh Joint Education Committee* resulted from our engagement with other project stakeholders in the planning process for the project.

Regarding e-language data, we were unable to collect data from Twitter or Facebook accounts because of issues of ownership restrictions, but website owners and blog authors cooperated generously and targets were exceeded. SMS messages proved more difficult to capture (much in the same way and for the same reasons as data collection through the app) but contributions through a dedicated WhatsApp number proved easier to elicit. Similarly, the collection of personal emails proved challenging while the collection of e-based workplace correspondence was easier to carry out. Contracts with BBC Cymru/Wales and S4C enabled the inclusion of samples of contemporary television and radio material, including podcasts. Notably, strong working relationships were developed with these institutions that led to them also supplying workplace emails, newsletters etc. for the written material.

Ethical approval for all aspects of data collection was obtained from all four of the universities involved in the project. Permission forms signed by participants included their

agreement for the collection of important metadata (e.g. age, gender, geographical location) necessary for the corpus. For a detailed discussion of the ethical considerations/challenges encountered when constructing CorCenCC, see Knight et al., 2021.

Transcription

Although transcription conventions for Welsh have been created for other projects (e.g. Deuchar et al., 2014), it was decided to create a bespoke set of conventions for the transcription of the CorCenCC data. These enabled us to fully reflect the whole spectrum of dialect/register variation captured in our speech data (making them more useful to academic researchers) as well as more accurately representing the speech of participants itself. Specifically, without an accurate representation of differences, it would not be possible to capture variations in Welsh, nor the distance between many spoken varieties and standard written Welsh. Transcribers were specifically instructed not to correct spoken patterns which might be considered non-standard or which included code-switching (i.e. switching between different languages during a communicative episode). Y Tiwtiadur (the pedagogic toolkit - see WP4) demanded that we should be able to identify *iaith anweddus* (inappropriate language) so that this could be systematically excluded from corpus applications used with children, for example, so transcribers were instructed to mark possible instances for review.

The recruitment of transcribers was an on-going challenge. As mentioned above, all prospective transcribers were given an initial test in which they had to transcribe a short piece according to the CorCenCC transcription conventions. A detailed overview, and rationale for, the decisions involved in the development of the transcription conventions for CorCenCC is available in Knight et al., 2021.

3.4. WP1: Key contributions

The main contribution of WP1 is the eleven million words of data that form the core of the corpus. In addition, the following range of resources, created as part of the data generation process, help achieve one of the stated aims of the CorCenCC project: to increase capacity and expand the interface between the Welsh language (and by extension other minoritised languages around the world) and the discipline of applied linguistics (including, in particular, corpus linguistics, sociolinguistics, and language planning and policy):

- A sampling frame for the creation of a general corpus of a minority language;
- A definition of 'inappropriate language' suitable for the context of a minority language where speakers are all bilinguals;
- A bespoke set of transcription conventions which can be applied to contemporary spoken Welsh;
- A team of Welsh-speaking research assistants who have been trained in the principles of corpus creation, who have had the opportunity to work with international experts across the world, and who are able to apply their skills to future projects.

The work from the WP1 team has been disseminated at a range of international and national conferences, and the publications (e.g. Knight et al., 2021) give more detail of the theoretical/methodological contributions of CorCenCC (see section 10 for references).

3.5. WP1: Applications and impact

The process of planning and implementing the core areas of work within WP1 offers a template for those researching other minoritised or minority languages. Feedback at international conferences confirmed awareness on the part of researchers elsewhere of the transferability of the CorCenCC methodology and processes to other language projects (e.g., corpus planning for Irish and Maltese, in recent conferences).

The corpus compiled under the auspices of WP1 will allow academic research into contemporary trends in the Welsh language. Examples of questions that might be addressed, given our data collection and transcription protocols, include:

- Are mutation patterns changing in contemporary Welsh and if so, which domains and genre are in the vanguard and rear-guard?
- How do the forms used in Welsh e-language compare to those in (various varieties of) spoken and written language?
- What does the corpus tell us about the current state of the 'traditional' Welsh geographical dialects and of new emerging varieties, including their social distribution?
- What words and phrases not previously formally recorded as part of Welsh are found in the corpus, and can any trends be identified?
- How prevalent is code-switching in spoken and e-language data and what appears to trigger it?

CorCenCC will reveal much about the current vibrancy and vitality of the Welsh language. We must be prepared for elements of these revelations to be welcomed and for other elements to be resisted by the wider Welsh-speaking community. Ultimately, should there be a community desire for it, CorCenCC can pave the way for an evidence-based discussion on what constitutes - or might constitute – 'standard Welsh' in the twenty-first century.

4. Work Package 2: Develop a part-of-speech tagset and use it to tag the WP1 data

4.1. WP2: Description

The purpose of WP2 was to operationalise a suitable means of identifying and labelling ('tagging') the parts of speech (e.g. noun, verb, and subtypes of such categories) featuring in the language collected in WP1, so that the corpus could be searched and analysed in the various ways that future users would require. The tools required were a part-of-speech tagger and a tagset. An existing resource, the Bangor Autoglosser (Donnelly and Deuchar, 2011), was used as the starting point, due to its demonstrable reliability and suitability for tagging Welsh. Under the guidance of WP2 leads Knight and project consultant Donnelly (a computational linguist who has worked closely on developing Welsh corpora), Neale (one of the project RAs) applied the Bangor Autoglosser to the emerging WP1 dataset, and identified where the tools needed adaptation and refinement—largely in relation to tagging spoken and

e-language sources, and the wider range of regional and genre variation. The modifications were applied in the latter stages of the project by Tovey-Walsh (a project PhD student), resulting in CorCenCC's own tagger and tagset, CyTag (see 4.3 below).

4.2. WP2: Objectives

The aims and objectives for WP2 were

- to construct and train the Welsh part-of-speech tagger
- to develop an appropriate tagset
- to tag all data

4.3. WP2: Achievements

CorCenCC's bespoke part-of-speech (POS) tagging software, CyTag, was developed in the first eighteen months of the project, and was publicly released in March 2018. Evaluations and applications of CyTag to date indicate that it is complete, robust, and working well. CyTag leverages open-source materials to help in the process of deciding on parts-of-speech. It works primarily by using the information found in Donnelly's Eurfa – the largest open-source, freely available dictionary for Welsh (Donnelly, 2013a) – to produce a list of possible tags for each word in a Welsh text. This is supported by specific lists of place names, first names and surnames extracted from Wikipedia data.

Once a list of words has been produced, a set of bespoke rules can then be applied to prune the list of possible tags for each word, based on the tags or features of its neighbouring words, until we arrive at the correct one. For example, the form 'yn' can mean 'in' but it also has two roles as a grammatical particle. In one role, it converts an adjective to an adverb (e.g. *yn dda* 'well', from *da*, 'good'). In the other, it is associated with the verb *bod* ('to be', one form of which is *mae* 'is') to introduce a verb, noun or adjective complement (e.g. *mae'r llyfr yn dda* 'the book is good'). As can be seen from these examples, *yn dda* can mean both 'good' and 'well', but in both cases *yn* would be classified in the same way. The tagger needed to distinguish *between* yn as a preposition and *yn* as a particle, and does so in the following way. In the sentence *mae Cymru yn wlad Geltaidd* ('Wales is a Celtic country') Cytag correctly tags 'yn' as a particle introducing a complement, because it is preceded by *mae,* and because 'wlad' is a soft mutation of 'gwlad' ('country'); we have a rule to select the complement particle tag for 'yn' if the following word is a soft- mutated noun. (When 'yn' means 'in' it is followed by a nasal mutation). The results of the evaluation process showed that CyTag was achieving an accuracy level of more than 95%, which is comparable with the best of the part-of-speech taggers for other languages. CyTag currently contains a text segmenter, a sentence splitter, a tokeniser, and the POS tagger itself. The website for CyTag is available at: http://cytag.corcencc.org.

The CyTag part-of-speech (POS) tagset contains 145 fine-grained, i.e. 'rich' tags, which are mapped into 13 categories compliant with the Expert Advisory Group on Language Engineering Standards (EAGLES, 1996). These tags include major syntactic categories (e.g. noun, article, preposition, verb and so on) as well as two categories representing 'unique' particles to Welsh and 'other' forms such as abbreviations, acronyms, symbols, digits etc. The full set of 145 tags covers Welsh morphology based on gender (masculine or feminine),

number (singular or plural), person (first person, third person, etc.) and tense (past, present, future, etc.). The tags themselves are encoded in Welsh.

The full tagset can be accessed at: https://cytag.corcencc.org/tagset?lang=en. See Neale et al., 2018 for a thorough technical overview of CyTag and an in-depth evaluation of its accuracy.

One important benefit of the CorCenCC approach to developing tagging software is its use of minimal, easily adaptable rules applied to existing knowledge and resources. This method is transferable to languages for which preannotated training data is scarce, making it valuable for capturing features of minority languages.

4.4. WP2: Key Contributions

The main contribution of the WP2 work was the creation of freely available software tools and linguistic resources which significantly extend the existing resource for Welsh language analysis and text mining. Specifically, we produced the following:

- The CorCenCC POS tagset (https://cytag.corcencc.org/tagset?lang=en), with:
  - 145 'rich' POS tags
  - 13 EAGLES-compliant 'basic' categories

  This tagset can either be adopted as a standard (i.e. set of conventions) and/or further enriched and/or adapted by future users when tagging Welsh language datasets.
- A gold-standard evaluation corpus. A gold-standard corpus is one that has been manually annotated and checked by multiple individuals. This effectively provides a model that can train and evaluate the automated (computerised) approach. This gold-standard evaluation corpus has also been released for other researchers to use in the development of their own tools. It comprises
  - 611 sentences
  - 14,876 tokens
- The CyTag website: (https://cytag.corcencc.org)
  - This website contains bilingual (Welsh and English) information about the tagger, including a working demo (see Figure 1a and 1b for screenshots in Welsh and English respectively). Users can access the tagger via this site and use it to tag their own data.
- CyTag on Github (https://github.com/CorCenCC/CyTag)
  - The open-source tagger has also been made available to all (i.e. available as free software, under the terms of version 3 of the GNU General Public License) via the GitHub website. Here, users can again download and use the tagger. They can also make improvements to the tagger and share updated versions with other, future, users.

The current version of CyTag includes the following improvements which were made by Tovey-Walsh:

- a revised version of the tokenizer
- an update to the way in which mutations are handled, improving its ability to capture mutation in shorter words and proper nouns

- inclusion of some words with a leading lower-case character in the lexicon (e.g. ‘cymraeg’)
- inclusion of high-frequency English words, enabling them to be identified by the tagger (as non-Welsh words), reducing the number of words tagged as ‘unknown’
- inclusion of Welsh determiners and indefinite pronouns in the tagset (e.g. ‘neb’ (=nobody), ‘pob’ (=every)) and tagging module, and added these rules to CyTag’s constraint grammar

Figure 1a. A screenshot of the CyTag online demo interface in Welsh (available at https://cytag.corcencc.org)

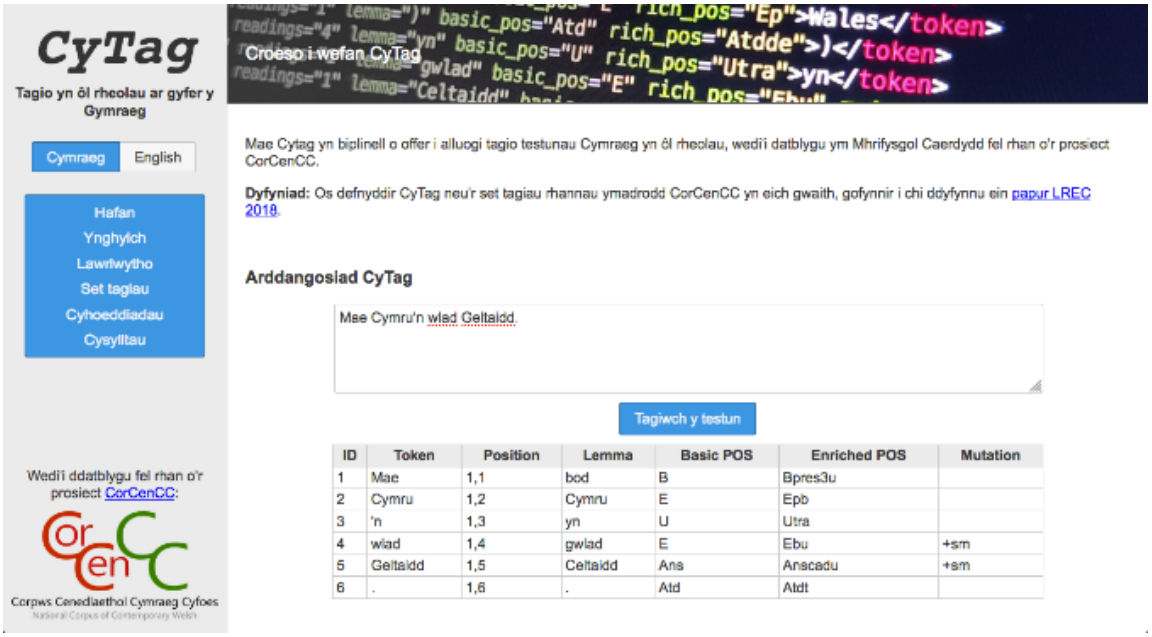
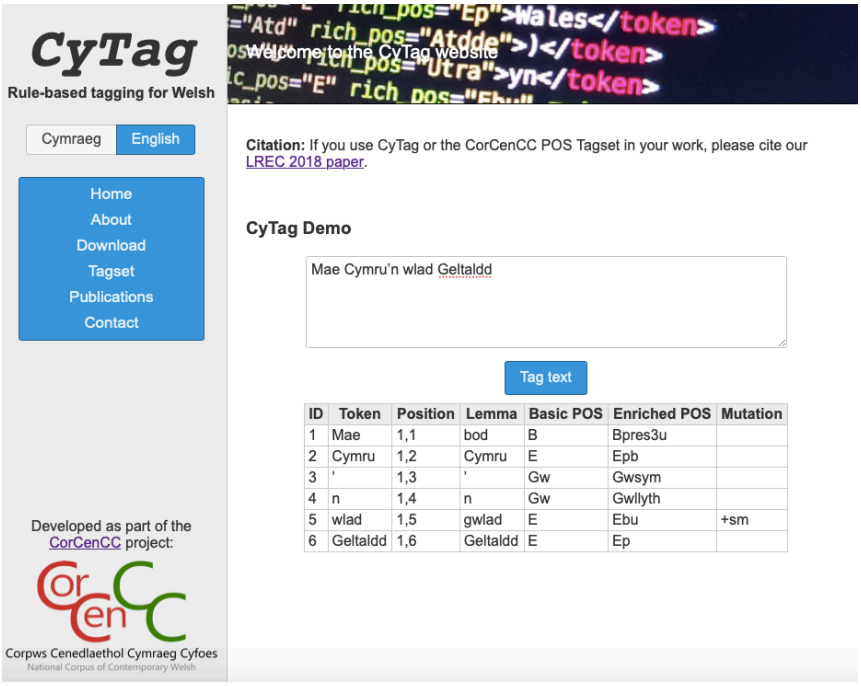


Figure 1b. A screenshot of the CyTag online demo interface in English (available at https://cytag.corcencc.org)



4.5. WP2: Applications and impact

Cytag is available in its own right as a tagger. This means it can be used by anyone who knows how to use a tagger, to tag their own data. Cytag has been pre-applied to the entire CorCenCC corpus, so that users do not need to know how to use a tagger to derive important information about the language. Such users might include researchers, language teachers, technology developers and lexicographers. CyTag can also be improved by other users in the future.

5. Work Package 3: Scope and develop a semantic tagger for Welsh and use it to semantically tag the WP1 data

5.1. WP3: Description

Work package 3 (WP3) was led by Rayson, with Piao acting as Research Associate until August 2018 (when he moved up to a fulltime lecturing position) at which point Ezeani joined the team. WP3's team drew on Welsh language expertise from across the whole CorCenCC project community where required. The major element of the research in WP3 was to design and develop the Welsh semantic annotation software system that would feed into associated linguistic resources.

5.2. WP3: Objectives

There were four aims and objectives for WP3:

- to design a novel semantic tagset for Welsh
- to develop the automatic annotation system
- to road-test crowdsourcing methods for semantic tagging
- to semantically tag all data

The research conducted in WP3 built on 30 years of work on automatic semantic analysis in corpus and computational linguistics in the UCREL research centre at Lancaster University. Of particular value for CorCenCC was the extensive work done on other languages (Finnish, Russian, Chinese, Dutch, Italian, Portuguese, Spanish, Malay).

As part of the CorCenCC project, we needed to re-evaluate the existing UCREL Semantic Analysis System (USAS) tagset in order to accommodate the special characteristics of Welsh, and the practical requirements of the pedagogic toolkit development (WP4) and the corpus end users (WP5). We also aimed to develop novel algorithms and methods to assign contextually appropriate semantic fields to Welsh lexical units, both single words and multiword expressions. Also, to complement the crowdsourcing method for corpus data collection in WP1, we aimed to pilot crowdsourcing methods for the assignment of semantic fields so as to extend the underlying semantic dictionaries and enable them to reflect the interpretations of Welsh speakers.

5.3. WP3: Achievements

The first task was to create a USAS tagset for Welsh language. This was achieved by drawing on methods developed in a previous AHRC-funded project, SAMUELS (AH/L010062/1),

and updating the Java framework for Welsh. The novel semantic analysis in Welsh involved the following steps.

A mapping across semantic fields from the existing multilingual framework was undertaken, so as to review its suitability for the Welsh language. The potential meanings assigned to tags were, therefore, first derived automatically by converting English dictionaries through bilingual dictionaries and small parallel corpora, and then checked by Welsh speakers from the CorCenCC team, and modified as required. The novel semantic tagset for Welsh was released in April 2016.

Our research on crowdsourcing entailed the creation of a method and experiments involving Amazon Mechanical Turk (AMT) users to investigate whether untrained non-professional Welsh speakers could create a reasonably high-quality semantic lexicon entry by assigning one or more suitable pre-specified semantic fields to a word or phrase from the corpus. This research was published in Rayson and Piao, 2017 (see section 10.2).

A part-of-speech tagger was needed for us to work on WP3, and in order to make progress with this before WP2 was complete, we temporarily used a pre-existing part-of-speech tagger, a component of the Welsh Natural Language Toolkit (WNLT) as part of the new semantic tagger for Welsh created in Java (CySemTag). Later, when the output from WP2 was available, we adopted CyTag. Initially a (SOAP) API website was made available on the UCREL GitHub account, enabling users to access the application. A new (REST) API was later made available at http://ucrel-api.lancaster.ac.uk/, to again increase accessibility to this application.

A Java framework for single word and multiword expression tagging was created by Piao, and using a variety of methods the linguistic resources were created. The final resources contain 143,287 single word entries and a collection of sample multiword entries, plus 329,800 inflectional forms from various corpora. Both web-based (using a SOAP API) and desktop GUI versions of the Welsh semantic tagger were created. In co-operation with the wider CorCenCC team, a gold standard corpus was manually checked for evaluation and tagger improvement purposes (details of the gold standard corpus are provided in 4.4).

To complete our research in this work package, Ezeani (RA) undertook a multi-task learning experiment to investigate whether state-of-the-art vector-based word embedding models for low-resource languages (in our case Welsh) could be used with neural network models for POS and semantic tagging. Our results showed that such an approach to tagging compared very well with the existing taggers (see Rayson and Piao, 2017 – section 10.2.) All our taggers and linguistic resources have been made available as open access with permissive licences.

5.4. WP3: Key contributions

A key contribution of the WP3 research is the freely available software tools and linguistic resources which augment the resource bank for Welsh language analysis and text mining. The CySemTag Java code is released on GitHub and has been incorporated into the Wmatrix (Rayson et al., 2004) corpus annotation and analysis system. This system is very widely used in corpus linguistics, and means that future researchers can use it for Welsh corpora. Overall the work in WP3 has extended the scope of research in corpus and computational linguistics in at least two ways. Firstly, it has demonstrated a method for effectively extending semantic

analysis techniques to the specific challenges of the Welsh language. Secondly, it has shown that crowdsourcing methods can be used to contribute to the development of such resources.

5.5. WP3: Applications and impact

The semantic tagger developed in WP3 has been applied to the full CorCenCC corpus, generating a rich interpretation of the data that will be of value to those interested in how contemporary Welsh is used to make meaning across genres, styles and mediums, and how language processes can be automated for new technologies. Furthermore, the underlying principles established in creating the CySemTag provide the basis for future extensions so that other low-resource languages can be analysed. Similarly, future researchers will be able to extend the multi-task learning experiments to further languages to investigate whether those languages, notwithstanding their differing grammatical frameworks, can also benefit. Incorporating the CySemTag into Wmatrix (Piao et al., 2018) extends its reach across the international research community, enabling the automatic content analysis of Welsh language corpora that might be collected by others in the future.

6. Work Package 4: Scope, design and construct an online pedagogic toolkit

6.1. WP4: Description

WP4 was led by Thomas and Fitzpatrick, working with project RAs Needs and J. Davies, and project lead Knight, with consultative advice provided by Stonelake (CI), Anthony (project consultant - designer and developer of AntConc) and Cobb (project consultant – developer of Compleat Lexical Tutor) and E. Davies (WJEC).

One field where corpora can be particularly informative is language learning/teaching. As teachers and learners become more adept with the use of technology, and as corpora continue to develop in size and application/functionality, corpus-informed learning is gaining a lot of ground both in classroom-based settings and in personal study. Corpora can be used to highlight the most common words, phrases and patterns in a language. They can show which words tend to go together, and which ones occur in which types of text (e.g. formal written texts, spoken conversations, professional e-mails, or personal text messages). Corpus users can search for specific words and see them in example sentences. Corpora therefore provide a rich source of language for the learner that demonstrates how their target language is actually used in practice, in various domains.

An innovative aspect of the CorCenCC project, led by the WP4 team, has been the development of a series of bespoke tools - Y Tiwtiadur - to be used within and outside Welsh language classes ranging from primary school to adult education. Together, these tools help demonstrate how Welsh is really used in four distinct corpus-based exercises that draw on data collected in WP1 and the taggers/tagsets developed in WP2 and WP3. The tools are:

- a Gap Filling (Cloze) tool allowing teachers (or learners, in self-study contexts) to delete words from any text in the corpus, at specified intervals to encourage or assess comprehension abilities and prediction strategies

- This tool allows users to create a gap fill task using texts from the CorCenCC corpus. The “Text type” option enables users to select particular genres of text (e.g. ‘blog’ or ‘book-fiction’ genres). The “Gap frequency” setting allows users to set the gap to appear as often as is desired, depending on how difficult the task needs to be (the recommended setting is every 7th - 9th word). Using the “Text length” option, users can choose to see a random sample of a text up to 100, 200, 300, 400, or 500 words long. On clicking “Start”, a new panel shows the gap fill task with the words that have been removed from the text appearing in a separate panel. To complete the task, users are required to choose words from the list and type them into the appropriate gaps in the text. When “Check” is clicked, the correctly placed words are highlighted in green, and the incorrectly placed words are highlighted in red. The availability of texts from the corpus enables teachers and learners to undertake this activity many times, with new opportunities for learning on each occasion

- a Word Profiler tool that enables the grading of texts by word frequency
  - This tool profiles a text selected or created by the user according to word frequency. Users are required to copy and paste a text into the “Input Text” area or type a text directly into the area. Clicking “Start” creates the profile, where each word is categorised according to its frequency level. In a separate panel, an explanation of the results is provided. The “Level”/“Frequency band” columns relate to the number of times a word appears in the 11-million-word CorCenCC corpus. Words in the “K1” (Top 1000) band are the 1000 most commonly used words in Welsh, according to CorCenCC. Typically, the more words the text has in the lower frequency bands (e.g. those in 3001-4000 (K4), 4001-5000 (K5) and >5001 (K6)), the more challenging it will be for the learner to comprehend. Learner-generated texts can also be profiled; typically, learners acquire more high-frequency words initially, and develop mastery of lower-frequency bands as proficiency develops (see e.g. Nation 2001). In the default setting, the tool will highlight words in levels K1 to K6+. Users can change the tool to highlight words that are not in these levels by clicking on the “Highlight non-level words” option. Words in the 5001+ band may include misspelled words and words from other languages, as well as vocabulary that is used infrequently in the corpus or is not captured in the corpus.
- a Word Identification tool testing learners' ability to guess a word in context
  - This tool displays multiple corpus extracts (concordance lines) that all contain a particular word. The word is blanked out, and the task is to identify the word that fits into all the gaps. There are options to select the “Frequency band” of the word (K1, K2 and K3), the specific “Word type” (e.g. noun, verb) and for the maximum number of sentences to be displayed. To generate the extracts, users click on “Start”. In this tool, the 'correct' responses given are those contained in CorCenCC; in some instances, a different response may also be plausible within the language more generally.
- a Word Task Creator tool that facilitates intensive work on a specified vocabulary item

- This tool generates multiple corpus extracts (concordance lines) from CorCenCC that all contain a target word specified by the user. Users type their target word into the "Word" entry box. They then select how many extracts they want to generate (maximum 20) using the "Maximum lines" option. If users want to specify the part-of-speech of the target word (e.g. noun, verb), they do so via the "Part of speech" option. The task is generated by clicking on "Start". This tool enables two types of learning activity. One is observation regarding the words surrounding the specified one. This assists with acquisition of grammatical structures and collocation patterns. The other is a more refined version of the Word Identification tool above, whereby a teacher can specify the actual word to be guessed, rather than only a word generated by the software from a set. To facilitate this second deployment of the tool, the target word is blanked out in the results table. Clicking on "Show" reveals the target word.

These tools enable learners to work with the language in multiple ways. For example, they can explore concordance patterns across various constructions (e.g. verb + preposition, adjective + preposition, conjunction (if, as) + following tense), and guess the missing word by inspecting and analysing other words used within its immediate co-textual and contextual environment. They can identify gaps in their vocabulary knowledge, and prioritise which words to learn next. Alongside the general query tools permissible within the CorCenCC corpus, *Y Tiwtiadur* offers a unique example of Data Driven Learning (Johns, 1991) in the form of inductive, direct-use, corpus-based pedagogy (Leńko-Szymańska and Boulton, 2015) that can help supplement Welsh language learning across the lifespan.

6.2. WP4: Objectives

The work conducted within WP4 responded to three objectives:

- to design and production of an online Welsh-based pedagogic toolkit (as described above) that works directly with the corpus data to support language teaching and learning.
- to produce frequency-based pedagogical word lists.

A key innovation of the CorCenCC project is that it integrates a corpus with an online pedagogic toolkit. Y Tiwtiadur works directly with the corpus data to support language learning and teaching by providing extensive opportunities to examine authentic Welsh language extracts. Whereas pedagogical corpus tools have hitherto been secondary adjuncts to pre-existing corpora, from the outset the design of CorCenCC included an educational interface to support Welsh language learning and teaching. Y Tiwtiadur was inspired by the online Compleat Lexical Tutor (Lextutor - https://www.lextutor.ca/ - Cobb, 2000) - one of the most high profile and frequently accessed (c.15000 users p/day) online data-driven language learning toolkits. The four tasks in Y Tiwtiadur are based on some of the most popular exercises featured in Lextutor.

A secondary innovation is that the toolkit was based on the concept of data-driven Learning (DDL - Johns, 1991), whereby 'learners inspect the evidence and look for patterns in the data from which they can form generalisations' (Thompson, 2005: 10). The four

pedagogical exercises function to enable and encourage learners (first and second language) to observe and extrapolate from Welsh language patterns in order to facilitate learning. This is crucial to learner autonomy (see Aston, 2001; Little, 2007). Rather than the teacher telling the student how the language works, the students are supported in working it out for themselves, utilising the concept of inductive (or 'discovery') learning, as advocated clearly within the constructivist approach to learning. This approach contrasts with deductive (or 'tutor-directed') learning, such as when learners are provided with structural rules. Integrating learning resources into the corpus, and constructing the corpus according to learner needs, enables learners to access relevant corpus data. The four pedagogical exercises built into CorCenCC facilitate focused attention to form, which is recognised as a key element of effective language learning.

Given that CorCenCC is the first corpus of its kind to draw on a wide variety of linguistic genres, linguistic forms, and linguistic representations of Welsh, a third innovation built into Y Tiwtiadur is the ability for learners and teachers/tutors to filter both the type and quantity of corpus data in generating their resulting outputs. In order that the learning facility is applicable to all ages, one key aspect of this third innovation has been the ability to filter out texts that are likely to include inappropriate content (such as expletives). Applicable to learners of all ages is the ability to distinguish between examples of language from different data types when looking specifically at uses of complex structures that may be used variably across speakers (e.g. Welsh mutation or some noun plural forms in texts of varying levels of formality). In this way, teachers and learners can manage the potential for confusion when more than one form is in use.

6.3. WP4: Achievements

The development of Y Tiwtiadur was user-informed from the outset. Tutors/teachers and learners, representing a wide range of Welsh language proficiency ability levels across the different language education sectors, were consulted at different time points to allow Y Tiwtiadur to be iteratively designed, tested and improved. The aims and objectives of WP4 were pursued across three main phases: (i) a consultation phase, (ii) a product development phase, and (iii) a showcasing phase.

(i) Consultation Phase

During the consultation phase, a questionnaire was piloted with a small number of practitioners in the field of Welsh language teaching to explore which resources learners and teachers already used and what they would ideally like to see developed as part of Y Tiwtiadur. Based on their feedback, a more detailed and targeted questionnaire was developed for a larger audience. This was distributed to Welsh teachers and tutors at two national conferences, and later shared in an online version. In all, 44 questionnaires were returned by Welsh teachers, trainers, lecturers and tutors from a wide variety of contexts – from those who teach at primary schools (Welsh-medium and English-medium) to those who teach adults – and 10 focus groups were conducted, accessing the views of 55 teachers/tutors and 14 Welsh for Adult learners.

In addition to the questionnaire, we met teachers and tutors face-to-face, to raise awareness of the corpus and Y Tiwtiadur, and to continue to collect views that would help

shape the development of the toolkit. The questionnaire responses along with follow-up focus group meetings helped us to identify the priorities for Y Tiwtiadur. The discussions we had were extremely useful, and the interchange of ideas that happened over the course of the focus groups and via the feedback obtained through the questionnaires enhanced our thinking about how to steer the work as we moved forward.

(ii) Product Development and (iii) Showcasing Phase

Following the consultation, the WP4 and WP5 teams collaborated to develop prototypes of the tool. This work was then developed further by J. Davies (supervised by Teahan) in collaboration with Anthony. There was an opportunity to demonstrate the work completed so far on Y Tiwtiadur at the annual conference of the National Centre for Learning Welsh in 2019, and tutors attending the workshops gave supportive and constructive feedback, including many helpful ideas with regard to usability. Their feedback informed the remainder of the development work. Positive insights included indications of how Y Tiwtiadur could be implemented with learners. Three main themes emerged:

1. Supporting the understanding of mutation at sentence vs. lexical level

A challenging feature of Welsh (and the other Celtic languages) is mutation – a morphophonological process whereby a phonological change is triggered in a closed set of word-initial consonants when certain words appear in particular syntactic contexts (Ball and Müller, 1992; Thomas and Mayr, 2010). For example, the initial 'c'/k/ sound in ***c****ath* /kɑθ/ 'cat' undergoes a process of lenition whereby /k/ is softened to /g/ - /gɑθ/. This change occurs after the definite article *y* "the", and after the numeral *dau* (masculine)/*dwy* (feminine) "two", for example. This phonological change is reflected in the written form, whereby *cath* is written as *gath*. In some cases, a word-initial consonant can be transformed in three ways, depending on the context (e.g. *cath* 'cat' /kɑθ/ (underlying form), *gath* /gɑθ/ (Soft Mutation), *nghath* /ŋ̊ɑθ/ (Nasal Mutation) and *chath* /χɑθ/ (Aspirate Mutation). These changeable forms can be problematic both in reading and writing, affecting vocabulary and literacy development/learning because they make it difficult to recognise words, and because the rules for the mutations are sometimes complicated. Corpus-driven examples can draw attention to form, and help learners identify where, when and how words change across contexts and the extent of any variation in the realisation of 'target' forms.

2. Supplementing gaps in dictionary support

Relating directly to mutation, as well as other changeable forms, many learners often fail to find words in paper and online dictionaries, either because they are looking for the mutated form (e.g. ***g****ath*) and not the underlying one (i.e. ***c****ath*) or because they mis-spelled or mis-represented the word in text. The corpus will allow learners to discover how a form of interest to them is typically used in various types of genre and medium.

3. Supporting learners in evaluating their writing

One key feature of Y Tiwtiadur is the option for learners and/or teachers/tutors to use the software to code a text of their own choosing for difficulty in terms of the frequency of forms used. This can be useful in determining the suitability of a text for a given learner or a group

of learners. In addition, learners can profile their own writing, so as to evaluate the extent of their own vocabulary knowledge/use.

In addition to these insights into how the tools might be implemented, stakeholders provided valuable reflections on the concerns they had about using such a resource in the classroom. Five key issues were raised, as highlighted below:

1. Modelling 'incorrect' language

A recurrent concern expressed among teachers and tutors from all educational contexts was that corpus data, unless 'corrected', might serve to model the very forms and expressions that teachers are trying to eliminate from their learners' attempts. Since a corpus expressly does not discriminate between what is considered correct and incorrect at a prescriptive level (other than by frequency of occurrence), teachers were advised to familiarise themselves in advance with the texts they would use in sessions, and identify anything that they wished to advise the learners about. This approach acknowledged that the relationship between descriptive and prescriptive approaches to language teaching is complex. It recognised that teachers do have a duty to inform learners about the forms that others will view as 'incorrect' (since such awareness is an aspect of knowledge about the language). At the same time, the manual approach would encourage teachers to question their own beliefs about what is and is not 'acceptable' as a target form, given the attested usage.

2. Offensive language

Many schoolteachers (from the Primary sector in particular) raised concerns about the possibility of accidentally accessing inappropriate content, focusing primarily on content that involved inappropriate or offensive words (such as expletives). Whilst it was beyond the scope of the current project to code for levels and types of offensiveness (which, in some instances, would not always be carried at the single word level), the corpus was tagged, at the text-level, for swear words and content of a particularly sensitive nature. While the statistical calculations that the pedagogic tools are based on reference the entire 11-million-word CorCenCC corpus, the four exercises of the pedagogic toolkit filter out these texts.

3. Complexity of Interface

One clear aim of CorCenCC was to develop a user-friendly corpus that could easily be transferred into education. For that to work, the program had to be fit-for-purpose and usable. This sentiment was echoed by some of the tutors and teachers we met. These views helped ensure that the four pedagogical exercises had a simple and clear interface that was intuitive to the user. At the same time, this interface mirrored the interface used by CorCenCC in order to ensure cohesion across the two and to instil confidence in teachers to progress from the pedagogic tools to a broader use of the corpus if desired.

4. Accessibility

Whilst the development of alternative platforms for running the corpus was beyond the scope of the present study, it was clear that in schools, in particular, where there is often limited access to computers, having an app that could be downloaded onto phones would be useful.

Schools already make use of existing Welsh language apps such as Duolingo and find that platform useful, so teachers enquired as to whether a similar platform could be developed for the current project. In sum, it was clear that a valuable future adaptation of Y Tiwtiadur will be developing an app for classroom use. As this was not one of the aims of the present project, and it would take considerable research and resource to complete, it has been noted as an important consideration for follow-on work.

5. Daunting number of examples in output

CorCenCC is a searchable corpus of over 11 million words. This means that some searches will result in huge outputs. For those who are not experienced with corpora and the types of outputs generated, the sheer volume of outputs produced can be overwhelming, and it could deter learners and teachers/tutors. For that reason, one of the innovative aspects of Y Tiwtiadur is that it allows teachers, tutors and learners the capacity to manage the amount of output from queries (as advocated in the Data Driven Learning approach). In addition to the ability to select topics or semantic categories when exploring text, the tool provides teachers/tutors and learners with the ability to limit text length when blanking out words in the Gap Fill exercise, provides a maximum output of 20 instances for the Word Identification and Word-in-Context exercises, etc. Together, these features increase the level of autonomy for the learner and teacher/tutor so that they are able to make the toolkit work for them.

6.4. WP4: Key contributions

Through the development of a pedagogical interface, led by the concept of data-driven learning and assessment, WP4 has contributed (i) a new pedagogical resource that is (ii) drawn from an online corpus of contemporary Welsh, of a kind that (iii) has never existed for the teaching of Welsh before, and that (iv) can serve as a model for similar work with other minority languages. It has made an invaluable contribution to language teaching and learning and, being open source, is available to support users in their inductive learning, irrespective of age, ability level and geographical location. The resource offers a new and unique opportunity for schools in Wales to embrace the concept of Data Driven Learning and to engage with and develop their own corpus-led pedagogies. In addition to this, teachers and learners, once introduced to the corpus via Y Tiwtiadur, might also feel confident enough to explore the corpus in other ways via the main CorCenCC query tools.

Various calls have been made in recent years for a corpus that can inform the delivery of Welsh (NFER, 2008: 48; Welsh Government 2013: 27, 71; Mac Giolla Chríost et al., 2012). CorCenCC, as a contemporary corpus of Welsh with an integrated pedagogic toolkit (Y Tiwtiadur), fulfils that need, informing curriculum writing, language assessment and language learning resources in the way that similar corpora do effectively for English (e.g. the Cambridge English Corpus (CEC) informs Cambridge English Language teaching resources; the British National Corpus (BNC) informs Pearson Longman's resources). An equivalent full and independent set of Welsh language teaching resources based on CorCenCC is a potential future direction of this work.

In line with the expectations laid out in the new *Curriculum for Wales: 2022*, such a resource will help develop learners' language awareness abilities and enrich their Welsh language skills in a naturalistic way, impacting ultimately on the Welsh Government's

Cymraeg 2050: Miliwn o Siaradwyr (Welsh Government, 2017) agenda, which aspires to a million Welsh speakers by 2050.

6.5. WP4: Applications and impact

The use of corpus data to support language learning in schools is a rapidly developing practice, but its effective implementation is under-developed and its effectiveness is under-researched. Within the New Curriculum for Wales 2022, children will be required to learn about the concepts of language, analyse linguistic nuances, and understand how these differ across languages. Corpus data is perfectly aligned with the promotion of such metalinguistic skills and knowledge, particularly within a bilingual context such as Wales. An important next step is to disseminate this free, on-line resource to schools in Wales. This will involve working closely with WJEC, curriculum designers and Welsh Government in identifying best practice for the use of CorCenCC in teaching and learning contexts across Wales, and modelling its implementation in other minority language contexts elsewhere. This work could lead to funded research projects evaluating the effectiveness of various applications of CorCenCC and Y Tiwtiadur with different types of learners with a view towards expanding its functionality where appropriate. For example, exercises could be adapted for a phone app or other technological platforms, increasing take-up and educational impact.

7. Work Package 5: Construct the infrastructure to host CorCenCC

7.1. WP5: Description

WP5 was concerned with the technical aspects of the construction of CorCenCC, building tools to support every stage of the process of corpus construction, from data collection (with a focus on the crowdsourcing app), through collation (via the data management tools), to querying and analysis (the web-based interface).

Spasić led WP5, working with Knight, Rayson, Piao and project RAs Neale and Muralidaran. Additional technical and consultative expertise was provided by Anthony (corpus linguist, educational technology specialist and creator of Antconc), Scannell (a computational scientist specialising in NLP, machine translation and minority languages) and Donnelly (a computational linguist who worked closely on developing previous Welsh corpora).

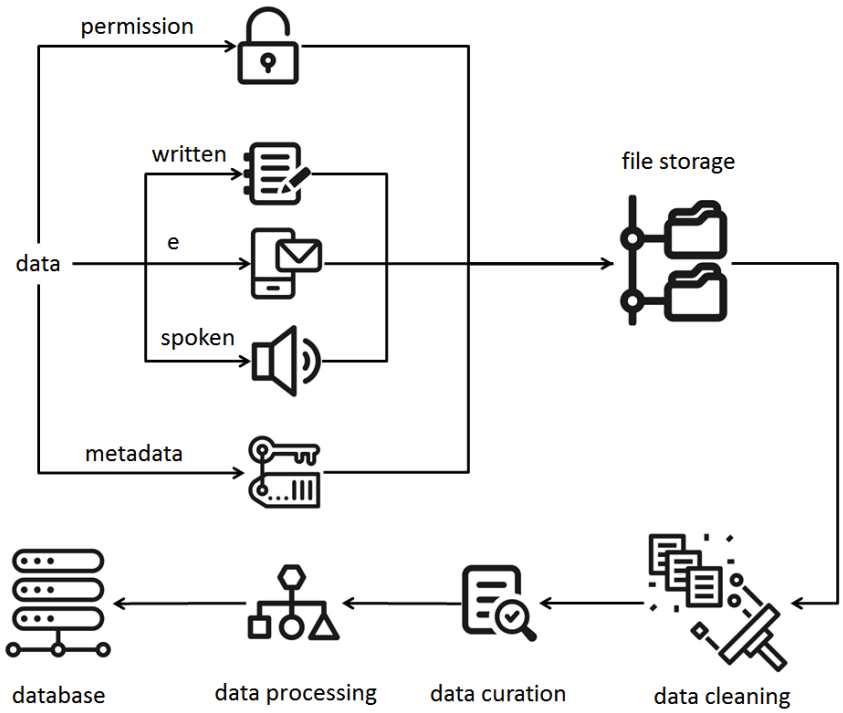
7.2. WP5: Objectives

WP5 aimed to develop a computational infrastructure to support the systematic collection and storage of this large quantity of text and analytic data together with a user-friendly interface to enable interaction with this data online. An important element was the design and construction of a repository system that would allow the adding of new data to the corpus over time, so that the maintenance of the corpus would be supported by its own users, and contributions to the corpus would be a social venture. A suite of corpus-analytic tools was developed on top of the repository to support functionalities that are typically integrated into contemporary corpora, such as KWIC (Key Word in Context) concordancers and collocation

tools, search and sort tools, word frequency lists, key word analysers and statistical testing facilities.

The data collection workflow is illustrated in Figure 2. The distinction between the three main language types (spoken, written and electronic (e)) is emphasised, as they required different processing before the data could be integrated into the corpus. As indicated in Figure 2, all relevant participant information and descriptive metadata was recorded at the time of data collection. Permissions to share the data in an online public resource were essential to the development of CorCenCC. These permissions were obtained from the relevant legal entities (e.g. the copyright owner; the speaker themselves) before the data was collected and locally stored. The raw data together with the corresponding permissions and metadata were deposited into a local file storage system. Subsequently, different data formats were standardised into plain text. Plain text can be processed automatically by natural language processing (NLP) tools, should another layer of linguistic metadata need to be added later; this helps future-proof the corpus, by enabling additional information to be added.

Figure 2. Data collection workflow in CorCenCC



7.3. WP5: Achievements

One of the key innovations of the CorCenCC project was redefining the design and construction of linguistic corpora, aligning methods more succinctly with the Web 2.0 age. To this end, steps were taken to construct and evaluate a system that enables 'live' user-generated spoken data collection via crowdsourcing. Crowdsourcing is a way to gather resources (in this case linguistic examples) from the general public, by making requests for volunteers to participate. We facilitated crowdsourcing by means of an app that can be run on

any Internet-enabled device; it ran in two forms: as a phone application and an interactive website. To maximise the potential user base, the mobile version of the app was implemented on both iOS (i.e. Apple) and Android platforms. The application made contributing to the corpus a very personal experience, giving users ownership and control of their own recordings.

All raw data (written, spoken and electronic) was stored systematically within a predefined folder structure, which corresponded to the sampling frame. From there, the data underwent the relevant cleaning and curation processes. To support collaborative multi-user access by the team members (from researchers to transcribers) across different sites (within and across the multiple institutions involved in the project), an online data management tool was developed on top of the file storage system. It provided a graphical user interface (GUI) that facilitated the uploading of raw data, indexing of the corresponding metadata and recording of subsequent data transformations, thus allowing the progress of all aspects of the corpus construction process to be monitored closely by all the researchers.

Once the texts had been converted to plain text format, they were marked up with layers of sociolinguistic metadata (e.g. source, genre, geographical origin) that would be used to query the data, and automatically tagged. As described in Section 4, the CorCenCC WP2 team developed CyTag (Neale et al., 2018) to achieve the tagging. CyTag is a suite of surface-level NLP tools for Welsh, based on the concept of constraint grammar (Karlsson, 1990, Karlsson et al., 1995). It supports text segmentation including sentence splitting and tokenisation as well as part-of-speech (POS) tagging and lemmatisation. It provides a bespoke solution for the basic linguistic pre-processing of Welsh including a tagset that is rich enough to capture idiosyncrasies of the language, notably in its spoken versions. To facilitate the semantic analysis of Welsh language data on a large scale, all pre-processed data was further marked up according to semantic categories using the CySemTagger developed in WP3 (see Section 5).

The corpus was stored and managed in a relational database where data could be accessed securely and concurrently by multiple users. To share the data online, we implemented a web-based interface to the database. The main reason for creating a bespoke interface rather than re-using an existing solution such as CQPweb (Hardie, 2012) was the requirement to tailor its functionality to the specific metadata of the CorCenCC corpus and its prospective users. To gather information about the user requirements, we used social media to survey current users of corpora. A total of 62 individuals responded, and their input identified the key functionality requirements.

An important consideration for the development of the corpus infrastructure was checking that it was fit for purpose. A group of corpus linguists evaluated the usability and functionality of the web-based interface. This process of evaluation involved a combination of questionnaires and talk-aloud exercises. Overall, the participants found the system useful in terms of meeting their information needs within the scope of their professional activities. The functionality was easy to understand without having to resort to help screen assistance. All participants agreed that they were likely to adopt the system and recommend it to other linguists.

7.4. WP5: Key contributions

The construction of a new corpus infrastructure is a major undertaking. The majority of corpus researchers use existing software to analyse the texts that they gather. Here, though, no such software existed, and it had to be built before the texts we had collected could be suitably indexed for entry into the corpus. Thus, we were simultaneously addressing several major challenges important for progressing corpus linguistic research.

Secondly, we necessarily designed from scratch much of the underlying computational infrastructure for tagging and analysing the language, given that Welsh has many features including grammatical distinctions, that do not transfer easily from languages with significant existing corpus resource (notably English), and substantial regional and register variation arising from the particular social history of the language. The key pillars of the infrastructure include a framework that supports metadata collection, an innovative mobile application designed to collect spoken data (utilising a crowdsourcing approach), a backend database that stores curated data and a web-based interface that allows users to query the data online. By using Welsh language tags, we have ensured that the corpus is not, and cannot be perceived as, an external (English) tool superimposed onto Welsh, but rather belongs to Wales and the Welsh language. Users will be encouraged by this means to buy wholeheartedly into the language as not only a source of information but also the medium through which it can be studied. At the same time, the availability of an additional English language interface will ensure an access point for the many whose interest in Welsh currently outstrips their facility with it, including the many thousands of learners of Welsh.

Thirdly, we have created tools that are freely available for others to adapt when creating their own corpora. We are particularly committed to supporting the building of corpora for other minority languages, and our user-driven model directly informs such projects by providing a template for corpus development in any other language.

Fourthly, by assembling an international team of experts, we have been able to deploy the latest technological innovations, and develop our own new ideas, lighting the pathway for future work in the Web 2.0 age. The crowdsourcing app is one of the first of its kind to be used for building a balanced corpus of natural language data by complementing more traditional methods of data collection, and successfully addressing a significant and persistent problem for the collection of high quality consented spoken data. Furthermore, we have demonstrated that it is possible to find the necessary manpower for transcribing such data and doing the essential manual tagging, even for a language with a relatively small cohort of fluent speakers.

7.5. WP5: Applications and impact

Though the computational infrastructure was developed for Welsh language collection, its design can be re-used to support corpus development in other minority or major language contexts, broadening the potential utility and impact of this work.

8. Summary of potential applications and impact

As described earlier in this document, every part of the project (as characterised by the work packages) has valuable applications that offer societal, economic and/or academic benefits. At a societal level, the corpus provides the opportunity to understand Welsh as a living language in use. In economic terms, the corpus offers scope to develop valuable new resources for Welsh learners and users, including potential for a corpus-based dictionary and a range of data-informed technological tools that might include language learning apps, predictive text production, word processing tools, machine translation, speech recognition and web search tools. Indirectly, the support of the corpus for these social and economic outcomes will promote the recognition of Welsh as a significant element of the UK and world linguistic landscape. The potential academic applications are broad and varied:

- Corpus linguistics: The innovative design of CorCenCC will inform and guide future corpus development and use in any language, via the provision of open access resources and the protocols for crowdsourcing approaches to data collection and analysis and for integrating research and teaching/learning functionalities.
- Language acquisition and bilingualism: The embedded data-driven learning facilities offer a unique opportunity for the investigation of autonomous learner behaviour and blended learning. CorCenCC incorporates a suite of frequency-based tools for research into patterns of acquisition and text and learner profiling.
- Sociolinguistics, dialectology and morphosyntactic analyses: CorCenCC extends the scope of available conversational data from the Siarad corpus (www.bangortalk.org.uk) to include speakers from a wider range of geographical regions. It will further inform us on the extent of language contact between English and Welsh. It will facilitate the exploration of sociolinguistic variation in patterns of lexical, grammatical, semantic, pragmatic and accent-based properties of language use within and across regions, by Welsh-user profile, thereby deepening the understanding of the social dynamics of Welsh as a living and reviving language.
- Language planning: CorCenCC provides base-line data with which to investigate language use in the context of policies relating to the use of Welsh in public administration, commerce and education.
- Lexicography: CorCenCC will be of direct relevance to the work of the team updating the University of Wales Dictionary of the Welsh Language (project partners), by providing primary evidence of lexical usage from attributable sources that can be included in future revisions of the dictionary.
- Computer and translation technology: As a tagged corpus, CorCenCC can be used for developing hybrid machine translation systems by feeding into a statistical machine translation engine, and facilitating the abstraction of grammatical rules from the corpus. Other areas of natural language processing (e.g. spelling correction, word prediction, assistive software for Welsh) will benefit from the more accurate and wide coverage language models that will be generated from the analysis of data from CorCenCC.

- Other research areas: CorCenCC offers new opportunities to academics engaged in stylistics and literary studies, media studies, psycholinguistics, pragmatics, business studies, health and medicine and psychology to extend their research into the Welsh language context.

CorCenCC is a freely available resource under an open licence which, when combined with the user-driven design and construction, will maximise its potential societal impact, informing the work and activities of current and future users of Welsh in a number of critical areas. Potential non-academic practitioner and professional domains include:

- Second language teaching and learning: As discussed in Section 6, reports on the teaching of Welsh for Adults (Mac Giolla Chríost et al., 2012; Welsh Government, 2013) have drawn attention to the need for a corpus of contemporary Welsh as a way to improve the efficacy of Welsh learning initiatives. CorCenCC functions to meet this need. By informing curriculum writing, language assessment and language learning resources as similar corpora do effectively in English (e.g. CEC and BNC), CorCenCC will facilitate data-driven learning, enhancing the effectiveness of teaching Welsh as a second language (compulsory in all schools in Wales up to the end of Key Stage 4). The anticipated impacts in the medium term are improved effectiveness in language learning; more nuanced awareness by teachers and learners of the inherent and natural variation in Welsh by region, genre and speaker-type; increased confidence in Welsh language speakers about the validity of their own usage patterns; and greater awareness about and pride in the Welsh language as a living and developing expression of Welshness.
- The Welsh Government and National Assembly of Wales (Language Policy): CorCenCC facilitates the realisation of action points in the Welsh Language Commissioner's strategy relating to digital content and applications, translation, terminology, language planning and research. These reflect the priorities of the Welsh Government (2014; 2017). As such, CorCenCC stands to impact the furthering of the integration of Welsh into everyday life as a language of governance, commerce and social interaction.
- The translation industry in Wales: CorCenCC outputs fit with the mid-term development of Microsoft Translate software. Preliminary research (Screen, 2014) shows that example-based machine translation alone can improve the productivity of human translators by up to 55%. By contributing to an eventual hybrid machine translation system, CorCenCC could further improve translation efficiency.
- The media in Wales: CorCenCC offers media companies the means to evaluate the linguistic nature of their output by measuring the difficulty of material (e.g. calculating the proportion of low frequency vocabulary) and establishing whose Welsh is being represented and underrepresented. By this means, CorCenCC could materially affect the accessibility and attractiveness of Welsh language programmes and media outputs to their target audiences, with consequent increased viewing/engagement figures, as well as supporting social equality.

- Welsh language publishers and lexicographers: CorCenCC provides the means to target content at audiences of different reading abilities and enhance the language tools available to authors for constructing graded readers. It will enable the commissioning of dictionaries of modern Welsh based on actual language use, thereby closing the recognised, and often problematic, gap between what Welsh users need and what they can find in reference sources. In turn this will build Welsh speakers' confidence in their linguistic abilities, making them more willing to use Welsh in a wider range of contexts.
- Language technology companies: a core requirement for companies using web-based and online social media data is a large, high-quality training corpus, and CorCenCC provides this. The CorCenCC dataset will, therefore, enable the development of a range of Welsh language technology resources that currently do not exist for the language.
- In the public domain: through its user-driven design, representatives of the likely future users of CorCenCC have been directly involved in the construction and design of the corpus which has ensured that it is user-friendly, accessible and appropriate to their needs. This approach functions to build on existing interest in Welsh language and heritage, and to foster community 'ownership' of the corpus. The potential long-term impact is a tangible change in perceptions of and attitudes to Welsh within and beyond Wales.

9. Associated projects and further funding

While CorCenCC itself was generously funded by the ESRC and AHRC, a range of satellite and associated research projects and other activities were funded from other sources. These are detailed below:

Date	Funder	Amount	Description [with PI]
Jan 2017	Cardiff University	£56,000	College of Arts, Humanities and Social Sciences (AHSS) funding received to support a three-year PhD scholarship for Vigneshwaran Muralidaran with a study entitled 'Using insights from construction grammar for usage-based parsing' [Knight and Spasić].
Feb 2017	British Council	£2,000	Funding to support the public launch of the CorCenCC project the Pierhead Building, Cardiff. [Knight]
Feb 2017	Swansea University	£1,000	RIAH - Research Institute for Arts and Humanities, Swansea University Funding received to support the CorCenCC project launch. [Fitzpatrick]
Feb 2017	Cardiff University	£1,500	School Research and Innovation Fund support received for the CorCenCC project launch. [Knight]
Oct 2017	Welsh Government	£24,992	Competitive commission from Welsh Government to provide a rapid evidence assessment of effective second language teaching approaches and methods. For more information see:

			https://tinyurl.com/ybtdsvfy [Fitzpatrick]
Jan 2018	Cymraeg 2050 2017-2018 Grant Scheme GC2050/17-18/20:	£19,964	Funding to construct a computational WordNet for Welsh. WordNet Cymru is lexical database in which words are grouped into sets of synonyms (synsets), which are then organised into a network of lexico-semantic relationships. To access the WordNet Cymru website, visit: http://corcencc.org/wncy/ [Spasić]
Jan 2018	Welsh Joint Education Committee (WJEC)	£1,968	Research grant (including intramural programme). Research grant to complete work on producing a B1 core vocabulary for Welsh for Adults (Canolradd level). For more information see: http://cronfa.swan.ac.uk/Record/cronfa48953 [Morris]
Mar 2018	Swansea University	£1,200	SPIN (Swansea paid internship) placement for data collection, transcription and interviewing of teachers/tutors 2017-18. Studentship for capacity building purposes. [Morris]
April 2018	Swansea University	£57,121	College of Arts and Humanities (COAH) funding to support a three-year PhD scholarship for Bethan Tovey-Walsh with a study entitled 'Purism and populism: The contested roles of code-switching and borrowing in minority language evolution'. Fees and maintenance paid [Morris and Fitzpatrick].
July 2018	Cardiff University	£2,100	CUROP (Cardiff University Research Opportunity) internal funding for a project entitled: 'Corpws Cenedlaethol Cymraeg Cyfoes: National Corpus of Contemporary Welsh - a focus on spoken data'. Studentship for capacity building purposes. [Knight]
July 2018	Cardiff University	£2,100	CUROP (Cardiff University Research Opportunity) internal funding for a project entitled: 'Corpws Cenedlaethol Cymraeg Cyfoes: National Corpus of Contemporary Welsh - semantic tagging and data annotation'. Studentship for capacity building purposes. [Knight]
Oct 2018	ESRC DTP Collaborative Studentship, Swansea University	£81,253	Welsh and Applied Linguistics: ESRC Wales Doctoral Training Partnership PhD Studentship entitled 'Strategic bilingualism: identifying optimal context for Welsh as a second language in the curriculum'. [Morris]
Jan 2019	Welsh Government	£20,000	Funding to support the development of a Welsh language Stemmer. [Spasić]
Aug 2019	Welsh Government	£90,000	Project entitled: 'Welsh language processing infrastructure: Welsh word embeddings'. Word embeddings are a type of word representation where words or phrases with a similar meaning are mapped to vectors of real numbers. The project focused on word embeddings for Welsh (primarily on creating a lexicon and Welsh word and term embeddings) and contributes to the Welsh Language Technology Action Plan's aim to 'promote Welsh language technology and coding

			resources to teachers and children and others'. [Spasić].
May 2020	Welsh Government	£90,000	Project entitled: 'Learning English-Welsh bilingual embeddings and applications in text categorisation'. This project aims to extend the results of the previous word embeddings project by creating cross-lingual representations of words in a joint embedding space for Welsh and English. [Knight]
		£451,198	

10. Summary of project outputs

10.1. Software tools

Name	Details	Link
CorCenCC's crowdsourcing app	Designed to allow Welsh speakers to record conversations between themselves and others across a range of contexts and to upload them, complete with ethically compliant consent from participants, for inclusion in the final corpus. Crowdsourced corpus data is a relatively new direction that complements more traditional language data collection methods, and is ideally suited to the positive community spirit that exists among speakers and learners of the Welsh language.	http://www.corcencc.org/app/ http://app.corcencc.org **Cite:** Knight, D., Loizides, F., Neale, S., Anthony, L. and Spasić, I. (2020). Developing computational infrastructure for the CorCenCC corpus – the National Corpus of Contemporary Welsh. *Language Resources and Evaluation (LREV).*
CyTag – Welsh Part of Speech Tagger	CyTag is an innovative Welsh tagger (complete with bespoke tagset) designed and constructed for the project. It is used in conjunction with the semantic tagger to tag all lexical items in the corpus.	http://cytag.corcencc.org **Cite:** Neale, S., Donnelly, K., Watkins, G. and Knight, D. (2018). Leveraging Lexical Resources and Constraint Grammar for Rule-Based Part-of-Speech Tagging in Welsh. Poster presented at the *LREC (Language Resources Evaluation) 2018 Conference,* May 2018, Miyazaki, Japan.
CySemTag Welsh Semantic Tagger Version 1	The Welsh Sematic Tagger applies corpus annotation automatically to Welsh language data.	http://ucrel.lancs.ac.uk/usas/ **Cite:** Piao, S., Rayson, P., Knight, D. and Watkins, G. (2018). Towards a Welsh Semantic Annotation System. *Proceedings of the LREC (Language Resources Evaluation) 2018 Conference,* May 2018, Miyazaki, Japan. Piao, S., Rayson, P., Knight, D., Watkins, G. and Donnelly, K. (2017). Towards a Welsh Semantic Tagger: Creating Lexicons for A

		Resource Poor Language. *Proceedings of The Corpus Linguistics 2017 Conference,* July 2017, University of Birmingham, Birmingham, UK.
CorCenCC's infrastructure and query tools	The CorCenCC query tools include the following functionalities: ▪ Simple query ▪ Complex query ▪ Frequency list generation ▪ Collocation analysis ▪ N-gram analysis ▪ Concordancing ▪ Keyword analysis	For tools and user guide, see: www.corcencc.org/explore **Cite:** Knight, D., Loizides, F., Neale, S., Anthony, L. and Spasić, I. (2020). Developing computational infrastructure for the CorCenCC corpus – the National Corpus of Contemporary Welsh. *Language Resources and Evaluation (LREV).*
Y Tiwtiadur	CorCenCC's pedagogic toolkit which is integrated with the main query tools. This includes the following teaching and learning tools: ▪ Gap-filling ▪ Vocab profiler ▪ Word identification ▪ Sentence-gap creator	For tools and user guide, see: www.corcencc.org/explore **Cite:** Davies, J., Thomas, E-M., Fitzpatrick, T., Needs, J., Anthony, L., Cobb, T. and Knight, D. (2020). *Y Tiwtiadur.* [Digital Resource]. Available at: www.corcencc.org/Y-Tiwtiadur

10.2. Publications (by reverse date order, with project team members in boldface):

1. **Knight, D., Morris, S., Arman, L., Needs, J.** and **Rees, M.** (2021a, in prep.). *Blueprints for minoritised language corpus design: a focus on CorCenCC.* London: Palgrave.
2. **Knight, D., Morris, S.** and **Fitzpatrick, T.** (2021b, in prep.). *Corpus Design and Construction in Minoritised Language Contexts: The National Corpus of Contemporary Welsh.* London: Palgrave.
3. **Knight, D.,** Loizides, F., **Neale, S., Anthony, L.** and **Spasić, I.** (2020). Developing computational infrastructure for the CorCenCC corpus – the National Corpus of Contemporary Welsh. *Language Resources and Evaluation (LREV).*
4. Corcoran, P., Palmer, G., **Arman, L., Knight, D.** and **Spasić, I.** (2020, accepted). Word Embeddings in Welsh. *Journal of Information Science*.
5. **Muralidaran, V., Knight, D.** and **Spasić, I**. (2020, accepted). A systematic review of unsupervised approaches to usage-based grammar induction. *Natural Language Engineering.*
6. **Spasić, I.**, Owen, D., **Knight, D.** and Arteniou, A. (2019). Data-driven terminology alignment in parallel corpora. *Proceedings of the Celtic Language Technology Workshop 2019,* Dublin, Ireland.
7. **Piao, S., Rayson, P., Knight, D.** and **Watkins, G**. (2018). Towards a Welsh Semantic Annotation System. *Proceedings of the LREC (Language Resources Evaluation) 2018 Conference,* May 2018, Miyazaki, Japan.
8. **Neale, S., Donnelly, K., Watkins, G.** and **Knight, D**. (2018). Leveraging Lexical Resources and Constraint Grammar for Rule-Based Part-of-Speech Tagging in Welsh.

Poster presented at the *LREC (Language Resources Evaluation) 2018 Conference,* May 2018, Miyazaki, Japan.

9. **Rayson, P.** (2018). Increasing Interoperability for Embedding Corpus Annotation Pipelines in Wmatrix and other corpus retrieval tools. Proceedings of the Challenges in the Management of Large Corpora workshop at the *LREC (Language Resources Evaluation) 2018 Conference,* May 2018, Miyazaki, Japan.
10. **Rayson, P**. and **Piao, S**. (2017). Creating and Validating Multilingual Semantic Representations for Six Languages: Expert versus Non-Expert Crowds. Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications held at the *European Chapter of the Association for Computational Linguistics 2017* (EACL) conference, April, Valencia.
11. **Piao, S., Rayson, P.,** Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R-M., **Knight, D.,** Křen, M., Löfberg, L., Nawab, R. M. A., Shafi, J., Teh, P-L., and Mudraya, O. (2016). Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. *Proceedings of the LREC (Language Resources Evaluation) 2016 Conference,* May 2016, Portorož, Slovenia.

10.3. Keynotes and invited talks

Research from the CorCenCC project has been presented at 17 keynotes and invited talks, and disseminated via 37 other conference papers delivered in 11 countries around the world. Details of these talks can be found on the main CorCenCC website (see: www.corcencc.org/outputs).

References

Aston, G. (2001) *Learning with Corpora*, Athelstan, Open Library.

Aston, G. and Burnard, L. (1997) *The BNC Handbook: Exploring the British National Corpus with SARA,* Edinburgh: Edinburgh University Press.

Ball, M. and Müller, N. (1992) *Mutation in Welsh,* Clevedon: Multilingual Matters.

Brabham, D. C. (2008) 'Crowdsourcing as a model for problem solving: An introduction and cases', *Convergence* 14**:** 75-90.

Carter, R. and McCarthy, M. (2004) 'Talking, creating: Interactional language, creativity, and context', *Applied Linguistics* 25**:** 62-88.

Cobb, T. (2000). *The compleat lexical tutor* [Online], available: http://www.lextutor.ca/ [Accessed 07/07/20].

Collins. (2020). *Collins Corpus online* [Online], available: https://collins.co.uk/pages/elt-cobuild-reference-the-collins-corpus [Accessed 07/07/20].

Cooper, Jones, D. and Prys (2019) 'Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology', *Information* 10**:** 247.

Cambridge University Press. (2020) *Cambridge English Corpus online* [Online], available: https://www.cambridge.org/us/cambridgeenglish/better-learning-insights/corpus [Accessed 07/07/20].

Deuchar, D., Webb-Davies, P. and Donnelly, K. (2018) *Building and Using the Siarad Corpus,* Amsterdam: John Benjamins.

Deuchar, M., Davies, P., Herring, J. R., Parafita Couto, M. and Carter, D. (2014) 'Building bilingual corpora: Welsh-English, Spanish-English and Spanish-Welsh', in Thomas,

E. M. and Mennen, I. (eds) *Advances in the Study of Bilingualism.* Bristol: Multilingual Matters.

Donnelly, K. (2013a) *Eurfa v3.0 - Free (GPL) Dictionary (incorporating Konjugator and Rhymer)* [Online], available: http://eurfa.org.uk [Accessed 07/07/20].

Donnelly, K. (2013b) *Kynulliad3: a corpus of 350,000 aligned Welsh and English sentences from the Third Assembly (2007-2011) of the National Assembly for Wales* [Online], available: http://cymraeg.org.uk/kynulliad3/ [Accessed 07/07/20].

Donnelly, K. and Deuchar, M. (2011) 'The Bangor Autoglosser: A multilingual tagger for conversational text', in Proceedings of *the Fourth International Conference on Internet Technologies and Applications (ITA11),* Wrexham, Wales. pp. 17-25.

Expert Advisory Group on Language Engineering Standards. (1996) *EAGLES guidelines* [Online], available: http://www.ilc.cnr.it/EAGLES/browse.html [Accessed 07/07/20].

Estellés-Arolas, E. and González-Ladrón-De-Guevara, F. (2012) 'Towards an integrated crowdsourcing definition', *Journal of Information Science* 38**:** 189-200.

Evas, J. and Williams, C. H. (1998) 'Community language regeneration: realising potential', in Proceedings of *the International Conference on Community Language Planning,* Cardiff, Welsh Language Board. pp. 1-13.

Hardie, A. (2012) 'CQPweb – combining power, flexibility and usability in a corpus analysis tool', *International Journal of Corpus Linguistics* 17**:** 380-409.

Hawtin, A. (2018) *The Written British National Corpus 2014: Design, compilation and analysis,* Unpublished PhD thesis: Lancaster University.

Johns, T. (1991) 'Should you be persuaded: Two samples of data-driven learning materials', *English Language Research Journal* 4: 1-16.

Karlsson, F. (1990) 'Constraint grammar as a framework for parsing running text', in Proceedings of *the 13th International Conference on Computational Linguistics (COLING),* Helsinki, Finland. pp. 168-173.

Karlsson, F., Voutilainen, A., Heikkilä, J. and Anttila, A. (1995) *Constraint grammar: A language-independent framework for parsing unrestricted text,* Berlin/New York: Mouton de Gruyter.

Knight, D., Adolphs, S. and Carter, R. (2013) 'Formality in digital discourse: a study of hedging in CANELC', in Romero-Trillo, J. (ed) *Yearbook of corpus linguistics and pragmatics,* Netherlands: Springer. pp. 131-152.

Leńko-Szymańska, A. and Boulton, A. (2015) *Multiple Affordances of Language Corpora in Data-driven Learning,* Amsterdam: John Benjamins.

Little, D. (2007) 'Language learner autonomy: Some fundamental considerations revisited', *Innovations in Language Learning and Teaching* 1: 14-29.

McEnery, T., Love, R., & Brezina, V. (2017) 'Compiling and analysing the Spoken British National Corpus 2014', *International Journal of Corpus Linguistics 22*(3): 311-318.

Mac Giolla Chríost, D., Carlin, P., Davies, S., Fitzpatrick, T., Jones, A. P., Heath-Davies, R., Marshall, J., Morris, S., Price, A., Vanderplank, R., Walter, C. and Wray, A. (2012). *Adnoddau, dulliau ac ymagweddau dysgu ac addysgu ym maes Cymraeg i Oedolion: astudiaeth ymchwil gynhwysfawr ac adolygiad beirniadol o'r ffordd ymlaen [Welsh for Adults teaching and learning approaches, methodologies and resources: a comprehensive research study and critical review of the way forward]*, Bedwas: Welsh Government.

Nation, I.S.P. (2001) *Learning Vocabulary in Another Language*, Cambridge: Cambridge University Press.

Neale, S., Donnelly, K., Watkins, G. and Knight, D. (2018) 'Leveraging lexical resources and constraint grammar for rule-based part-of-speech tagging in Welsh' in Proceedings of

the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan. pp. 3946-3954.

NFER (2008) *Ymchwil i'r Cwrs Dwys ar gyfer Cymraeg i Oedolion*, Swansea: National Foundation for Educational Research.

ONS (2011) *DC2612WA – Ability to speak Welsh by occupation* [Online], Office for National Statistics, Durham: Nomis. Available: www.nomisweb.co.uk/census/2011/dc2612wa [Accessed 07/07/20].

Piao, S., Rayson, P., Knight, D. and Watkins, G. (2018) 'Towards a Welsh semantic annotation system' in Proceedings of *the Eleventh International Conference on Language Resources and Evaluation (LREC),* Miyazaki, Japan. pp. 980-985.

Rayson, P., Archer, D., Piao, S. and McEnery, T. (2004) 'The UCREL semantic analysis system', in Proceedings of *the Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP tasks at the 4th International Conference on Language Resources and Evaluation (LREC),* Lisbon, Portugal. pp. 1-6.

Scannell, K. (2007) 'The Crúbadán Project: Corpus building for under-resourced languages' in *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*. pp. 1-10.

Scannell, K. (2012) *Kevin Scannell's website* [Online], available: http://borel.slu.edu/ [Accessed 07/07/20].

Sinclair, J. (2005) 'Corpus and text - basic principles', in Wynne, M. (ed) *Developing Linguistic Corpora: a Guide to Good Practice,* Oxford: Oxbow Books. pp. 1-16.

Thomas, E. M. and Mayr, R. (2010) 'Children's acquisition of Welsh in a bilingual setting: a psycholinguistic perspective, in Morris, D. (ed) *Welsh in the 21st Century,* Cardiff: Cardiff University Press.

Thompson, P. (2005) 'Spoken Language Corpora' in Wynne, M. (ed) *Developing Linguistic Corpora: a Guide to Good Practice,* Oxford: Oxbow Books. pp. 59-70.

Welsh Government (2013) *Codi golygon: adolygiad o Gymraeg i Oedolion. Adroddiad ac argymhellion [Raising our sights: review of Welsh for Adults. Report and recommendations]*, Bedwas: Welsh Government.

Welsh Government (2014) *A Living Language: a language for living - Moving forward*, Cardiff: Welsh Government

Welsh Government (2017) *Cymraeg 2050: A million Welsh speakers Action Plan 2019-20*, Cardiff: Welsh Government.

Wilson, A. (2002) *The Language Engineering Resources for the Indigenous Minority Languages of the British Isles and Ireland Project* [Online], available: https://www.lancaster.ac.uk/fass/projects/biml/default.htm [Accessed 07/07/20].