Lang Resources & Evaluation
https://doi.org/10.1007/s10579-020-09501-9

ORIGINAL PAPER

Developing computational infrastructure for the CorCenCC corpus: The National Corpus of Contemporary Welsh

Dawn Knight[1] · **Fernando Loizides**[2] · **Steven Neale**[1] · **Laurence Anthony**[3] · **Irena Spasić**[2]

Abstract CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes—National Corpus of Contemporary Welsh) is the first comprehensive corpus of Welsh designed to be reflective of language use across communication types, genres, speakers, language varieties (regional and social) and contexts. This article focuses on the computational infrastructure that we have designed to support data collection for CorCenCC, and the subsequent uses of the corpus which include lexicography, pedagogical research and corpus analysis. A grass-roots approach to design has been adopted, that has adapted and extended previous corpus-building and introduced new features as required for this specific context and 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. A usability study was conducted to evaluate the user facing tools and to suggest directions for future improvements. Though the 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.

Keywords Language resources · Natural language processing · Data modelling · Information retrieval · Web interfaces · Usability testing

✉ Dawn Knight
KnightD5@cardiff.ac.uk

[1] School of English, Communication and Language Sciences, Cardiff University, Cardiff CF10 3EU, UK

[2] School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, UK

[3] Faculty of Science and Engineering, Waseda University, Tokyo, Japan

Published online: 29 July 2020

1 Introduction

Corpora, and their associated concordancing software (enabling the investigation of language patterns across corpus data, the creation of frequency lists, etc.), help to generate empirically based, objective analyses that are "focused not simply on providing a formal description of language but on describing the use of language as a communicative tool" (Meyer 2002). Corpus-based methods have revolutionised the investigation of language and they continue to increase the significance and reach of their applications, both within and beyond language description and teaching and learning. The potential for corpus investigation in any given language is of course dependent on the availability of a comprehensive and contemporary resource. UK-based researchers were amongst the pioneers of the field of corpus linguistics, who originally focused primarily on the English language, although their methodologies are now used widely beyond English. However, Welsh, the second most used language in the UK (Office for National Statistics 2011), lags behind other European languages in terms of corpus provision, with no resource to match, inter alia, the German reference corpus DeReKo (Kupietz et al. 2018), the Czech National Corpus (Kucera 2002), the Croatian National Corpus (Tadić 2002), the ZT Corpus of Basque (Areta et al. 2007) and the CUCWeb corpus of Catalan (Boleda et al. 2006).

While a number of Welsh language corpora do exist, the majority contain language data from one medium of communication (e.g. the spoken Siarad corpus (ESRC Centre for Research on Bilingualism 2020; Deuchar et al. 2014) and the written Cronfa Electroneg o Gymraeg (CEG) (Ellis et al. 2001)) and/or are relatively small in scale. In addition to this, several are context-specific and/or specialist in nature (genre specific), having been designed specifically to support research in areas such as code-switching, bilingualism or translation (e.g. the Welsh speech database (Williams 1999)). Others, such as the e-language corpus from the Crúbadán project (Scannell 2007), which are extensive in size, contain data compiled in an unstructured way (using web crawlers), without a predefined balance across text types, themes and genres or across contributor profiles; furthermore, permission to (re)use the data in a corpus was not sought from individual contributors.

Corpws Cenedlaethol Cymraeg Cyfoes (CorCenCC) (D Knight et al. 2017; Knight et al. 2020), The National Corpus of Contemporary Welsh, was designed to overcome these shortcomings. The 10 million-word CorCenCC is the first comprehensive corpus of Welsh that is reflective of language use across different communication types (spoken, written, e-language), genres, language varieties (both regional and social), thematic contexts and contributors (i.e. from a range of different demographic profiles). Drawing on the latest international expertise, and developing a range of innovative techniques, the corpus was compiled in a principled way with full permissions obtained from the contributing legal entities, including individuals, groups and organisations. The corpus was designed to provide resources for the Welsh language that can be used in language technology (speech

recognition, predictive text, etc.), pedagogy, lexicography and academic research contexts, amongst others.

This paper maps out the technical aspects of the construction of CorCenCC, providing a description and evaluation of the bespoke tools that were built 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). As such, it represents a model that will support future corpus design, in sharing a template for corpus development in any language.

2 Related work

The construction of written corpora is relatively straightforward, since text is commonly available in computer-readable (digital) formats that can be scraped, processed and uploaded for use at the click of a button. As a result, the corpus linguistics field has a number of very large English corpora, e.g. the COCA (520 M words) (Davies 2010), GloWbE (1.9B words) (Davies and Fuchs 2015), and enTenTen (19B words) (Jakubíček et al. 2013), all featuring primarily written language, although some spoken elements are often included. In addition, a number of corpora now include samples of language used in social media and other web contexts alongside more traditional written and (transcribed) spoken language. One example is the BNC 2014 corpus (Love 2020) which, though primarily structured to be comparable with the original 1994 version of the BNC (British National Corpus)—see Aston and Burnard (1998)), also includes genres (including digital) that did not exist at the time the original corpus was constructed.

The development of comparably substantial/large-scale spoken corpora has lagged behind, mainly due to the time-consuming nature of recording, transcribing and processing spoken content. Despite the fact that there is an ever-increasing availability of, for example, speech-to-text utilities which assist with the automation of processing audio recordings, the quality of these tools falls short of needs of linguists. This is because such tools fail to fully and reliably capture and account for the subtle intricacies of speech (e.g. overlaps, false starts, hesitations and backchannelling behaviours). Until this is remedied, more costly, manual, approaches to spoken data preparation continue to be required.

As discussed by Adolphs et al. (2020), the majority of the recently developed spoken corpora consist of material such as transcripts of radio talk shows and television news, since their collection can be easily operationalised as text-based scripts of (at least some of) the talk already exist. The nature of this spoken discourse is typically described as unscripted; however, it is certainly constrained, e.g. talk show radio has certain expectations about how the host will moderate the discussion. While the scripted/constrained oral content in these spoken corpora is facilitative in the exploration of spoken discourse, it can be challenged in terms of the extent to which it can be considered evidence of natural language use. In short, such material is by no means a direct or reliable substitute for spontaneous, unscripted oral discourse.

Even with the collection of scripted/constrained spoken discourse, the largest spoken corpora remain small in size in comparison to their written counterparts. The most notable spoken corpora include Michigan Corpus of Academic Spoken English (MICASE) at circa 1.8 M words (Simpson-Vlach and Leicher 2006), the Cambridge and Nottingham corpus (5 M words) (Carter and McCarthy 1997), the Database for Spoken German (8 M words) (Schmidt 2014) and the spoken component of the BNC, both versions of which contain 10 M words of spoken context-governed and/or informal English. Arguably, the most extensive spoken corpus in existence is the language 'banks' collectively available through the TalkBank system (MacWhinney 2000). TalkBank is an online corpus resource which supports the construction, sharing and collaborative analysis of individual clinical, child language, multilingual, conversational and other spoken datasets, through the utility of standardised transcription and mark-up conventions and coding/analysis tools. Collectively, this resource brings together hundreds of researchers working in over 34 languages.

Given the restricted timespan and scale of the initial CorCenCC project (3.5 years, with a pre-specified funding pot), the focus was creating an initial dataset of 10 M words of written and spoken contemporary Welsh language, taking a user-friendly and replicable approach to enable the exploration and extension of the dataset beyond the lifespan of the project. While 10 M words is far fewer than its English language contemporaries can boast, CorCenCC is nevertheless a significant milestone in corpus development. It is the first comprehensive corpus of Welsh, which covers language use across different communication types (spoken, written and electronic ('e')), genres, language varieties (regional and social), and contexts, and in reflecting the breadth of the population of 562,000 Welsh speakers in the UK.

A user-driven perspective of existing commonalities in corpus-building tools drove the design of those developed for CorCenCC (as examined in Sect. 3.7). The majority of existing corpora either provide their bespoke query interfaces for web-use, such as Sketch Engine (Kilgarriff et al. 2014) or else the corpora can be downloaded and analysed offline using concordancing software such as AntConc (Anthony 2014). While there is no one strict 'standard' approach to corpus analysis (as it depends on the specific research questions at hand), there are commonalities in the suite of analytic tools available, which commonly include search, sort, word list, concordance, keyword and n-gram functionalities. These preexisting technological resources were a valuable starting point for the present project.

3 Methods

3.1 Data collection workflow

The first objective of the CorCenCC project was to create a corpus that was not only large in scale, but also 'representative of a wider range of language samples/types, to enable the linguist to make observations of language-in-use from a multitude of different perspectives' (Knight 2011; Leech 2014; Atkins et al. 1992; Biber 1993). Therefore, we implemented a principled approach to data collection (Rees et al. 2017), addressing Sinclair's (Sinclair 2005) call for corpora to provide, amongst

Table 1 Examples of the sources of content included in the CorCenCC corpus

Data type	Potential sources/text-types
Spoken	Televised interviews and TV chat shows (BBC); BBC radio shows; political speeches
	Conversations with friends; conversations with family members
	Workplace Welsh; service encounters; phone calls; formal and informal interaction at the National Eisteddfod
	Welsh learner discourse; Primary, secondary, tertiary and adult classroom interaction
Written	Books; papurau bro (i.e. local community Welsh language monthlies); political documents; stories
	Welsh learner writing (at various ages and/or levels of proficiency)
	Letters and diaries; academic essays; academic textbooks; magazines; adverts, flyers/information leaflets; formal letters
	Signs (e.g. public information signs, road signs)
Electronic	Website content; blogs
	Emails
	Text messages

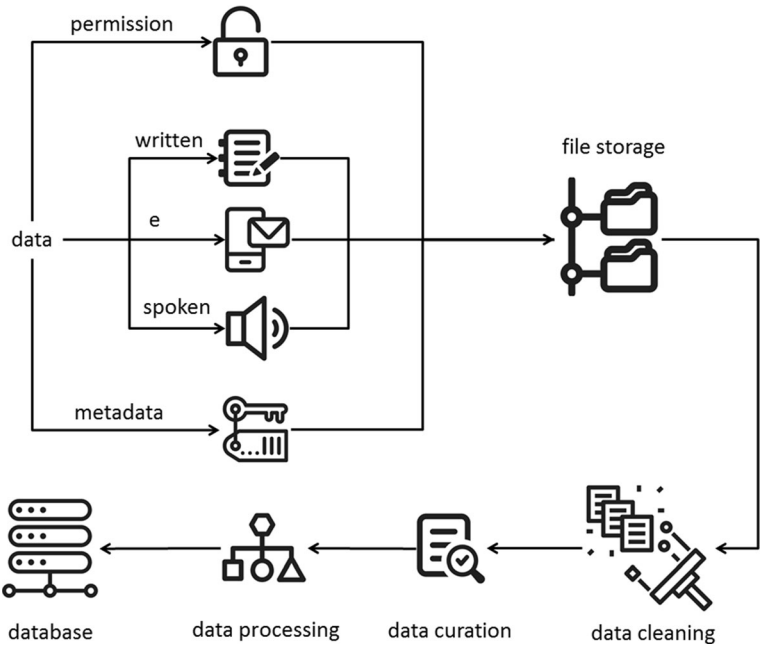


Fig. 1 Data collection workflow

other things, 'representative' and 'balanced' samples of 'fully documented' language data, along with further relevant metadata.

The design of CorCenCC's sampling frame includes data from different themes, topics and types, informed by the thematic groupings and discourse categorisations of existing corpora including Cambridge and Nottingham Corpus of Discourse in English (CANCODE) and Cambridge and Nottingham e-language Corpus (CANELC) (see (Carter and McCarthy 2004) and (Knight et al. 2013)). Thus, the corpus includes data from formal contexts (e.g. political documents, televised interviews and formal letters) to less formal ones (e.g. diaries, phone calls and text messages)—see Table 1 for examples of the types of Welsh-language data sourced for CorCenCC. The distinction between the three main language types is emphasised in Fig. 1, as they require different types of processing before the data can be integrated into the corpus. In terms of individual contributors to the corpus, we collected data from a wide range of ages, occupations and genders located in different geographical regions of Wales. Learners of Welsh constitute a significant proportion of the Welsh speaking population, and the corpus also reflects this through its inclusion of learner data. To monitor the balance of the corpus as well as to support future studies of the corresponding sublanguages, all relevant metadata were recorded at the time of data collection (see Fig. 1).

With regard to permissions, a particularly extensive range was sought, because of the planned destination of the corpus. The CorCenCC corpus is available under an open licence to academic researchers and community users of Welsh, and also to the

Table 2 Categories of metadata

Category	Definition	Domain	Examples	Language		
				S	W	E
Type	The text-type of written, spoken or digitally-based communication.	Taxonomy	Letter → Professional	x	x	x
Genre	A type of textual material distinguished by the style and/or contents.	Taxonomy	Factual → Tourism and travel		x	x
Scripted	The content of a conversation was scripted?	Boolean	True, false	x		
Translated	Text was translated from another language?	Boolean	True, false		x	x
Source	A legal entity that contributed the data with permission to use.	Controlled vocabulary	BBC radio cymru	x	x	x
Target audience	A particular group at which a text document is aimed.	Controlled vocabulary	Adults		x	
Context	A communicative situation that influences the discourse structure.	Taxonomy	Media → TV → News and weather	x		
Participants	Social relations between participants in a conversation.	Taxonomy	Family → Siblings	x		
Location type	A physical environment where a conversation took place.	Taxonomy	Public spaces → Shops	x		
Place name	The name of a geographical region where the recording took place.	Gazetteer	Caerphilly	x		
Demographics	A set of classifiable characteristics of an individual person.	Faceted	See Table 3	x		
Language ability	A set of properties that describe one's Welsh language ability.	Faceted	See Table 4	x		

Table 3 Categories of demographic metadata

Category	Definition	Domain	Example
First name	A person's given name	Free text	Rhys
Last name	Family name of person	Free text	Jones
Gender	Subject self-identification regarding their masculinity and femininity	Controlled vocabulary	female
Birth year	The year in which a person was born	Integer	1984
Residence	An individuals' current place of residence	Gazetteer	Cardiff
Occupation	A regular employment activity performed for payment	Controlled vocabulary	Lower supervisory and technical occupations
Employment status	Quantification of an individual's occupation in terms of employment	Controlled vocabulary	Full-time student
Legal entity	A person, company or organisation that has legal rights and obligations	Controlled vocabulary	Person

Table 4 Categories of metadata related to a speaker's Welsh language ability

Category	Definition	Domain	Examples
Dialect	The regions of Wales that have most influenced a speaker's dialect	Gazetteer	North West
Learning environment	The environment where learning of Welsh occurred. Not limited to educational settings	Controlled vocabulary	Welsh-medium/bilingual secondary school
Proficiency	A self-reported measurement of how well a person has mastered the language	5-point Likert scale	I can communicate basic information in everyday situations
Learner	Self-identified as a learner of Welsh?	Boolean	True, false
Level	For learners of Welsh, the level of the most recent Welsh language course taken	Free text	Beginner

wider European language technology community. While many existing large-scale and national corpora can be accessed easily, few are available under an open licence due to copyright restrictions and permissions. Specifically, many are owned by publishers. 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 before the data were collected and locally stored.

The raw data together with the corresponding permissions and metadata were deposited into a local file storage system (see Fig. 1). Subsequently, different data formats were standardised into plain text. Plain text can be processed automatically by a suite of natural language processing (NLP) tools (see Sect. 3.5) to add another layer of linguistic metadata; this will facilitate future studies of the Welsh language. Some operations during the data format conversion were manual (e.g. transcription, anonymisation and quality control) while others were automated (e.g. removal of non-textual content). The following sections describe the data collection and management steps in more detail.

3.2 Metadata collection framework

Metadata are typically described as data about data (Duval et al. 2002). In other words, they consist of structured information that describes, explains, locates or otherwise makes it easier to retrieve the underlying data. In this section we focus specifically on descriptive metadata, which are used to facilitate searches for data using descriptors that qualify their content in CorCenCC. Structural metadata, which describe how data components are organised in CorCenCC, will be described separately, in relation to data storage. As Table 2 illustrates, different categories of metadata are recorded for each of the different types (i.e. sources) of linguistic data included in CorCenCC.

3.3 Spoken data collection

The collection of written and e-language data focused on readily accessible data. However, spoken data of spontaneous or unscripted nature are not abundant and required creation of effective means of eliciting new data from Welsh speakers. One of the key innovations of the CorCenCC project was to redefine the design and construction of linguistic corpora, aligning methods more succinctly with the Web 2.0 age. Steps toward achieving this aim were taken in constructing and evaluating a system which enables 'live' user-generated spoken data collection via crowd-sourcing, an "online, distributed problem-solving and production model" (Brabham 2008) involving "Internet-based collaborative activity, such as co-creation and user innovation" (Estellés-Arolas and González-Ladrón-De-Guevara 2012). Application of crowdsourcing methods in linguistics is still generally at a relatively early stage of development, but crowdsourcing methods have been used extensively in other fields such as computer science (including NLP), and are increasingly being applied in the arts, humanities and social sciences, e.g. (Weinberger 2020; Mozilla 2020). Crowdsourcing methods were also recently used in the development of the spoken BNC 2014 corpus (McEnery et al. 2017), with potential contributors being

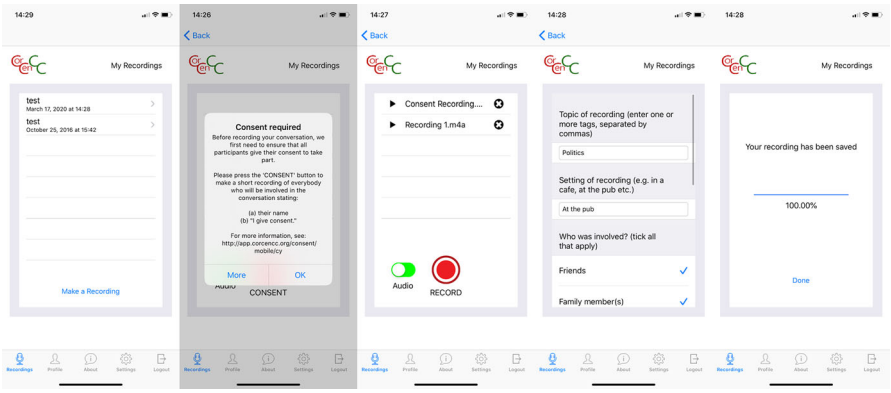


Fig. 2 The screenshot of a user's dialogue on the crowdsourcing app

remunerated for recording their own conversations and sending them to Cambridge University Press for further processing and integration into the corpus. Crowdsourcing has also been used to collect Welsh spoken data with the aim of supporting speech recognition rather than corpus linguistics studies (Prys and Jones 2018). This project focused primarily on collecting voice recordings of scripted material unlike CorCenCC, which focuses exclusively on spontaneous, unscripted discourse.

We facilitated crowdsourcing by an application that can be run on any Internet-enabled device, either via live, a mobile app, or by uploading pre-recorded files via an interactive website (Neale et al. 2017). The crowdsourcing application is one of the first of its kind to be used to complement more traditional methods of data collection in pursuit of a balanced corpus of natural language data. The application allows users to collect and upload their spoken language data complete with the corresponding metadata and permissions to use them. To maximise the potential user base, the mobile version of the app was implemented on both iOS and Android platforms. Figure 2 provides a screenshot that illustrates one of the use cases. The application made contributing to the corpus a much more personal experience, giving users ownership and control of their own recordings. No financial incentives were offered to the users who decided to donate their data to the project.

3.4 Data management

All raw data (written and spoken) were stored systematically into a predefined folder structure, which corresponded to the sampling frame. From there, they underwent the relevant cleaning and curation processes for that data type. Data cleaning involves data format conversion (e.g. from PDF or Microsoft Word to plain text) and basic pre-processing such, as the removal of non-textual content (e.g. URLs, menus, page navigation links, etc.). Data curation involved the transcription of the spoken data and anonymisation. Since there were several paid transcribers, there was also quality control. Transcriptions were subjected to a broad quality check on receipt to ensure that transcription conventions had been followed and that

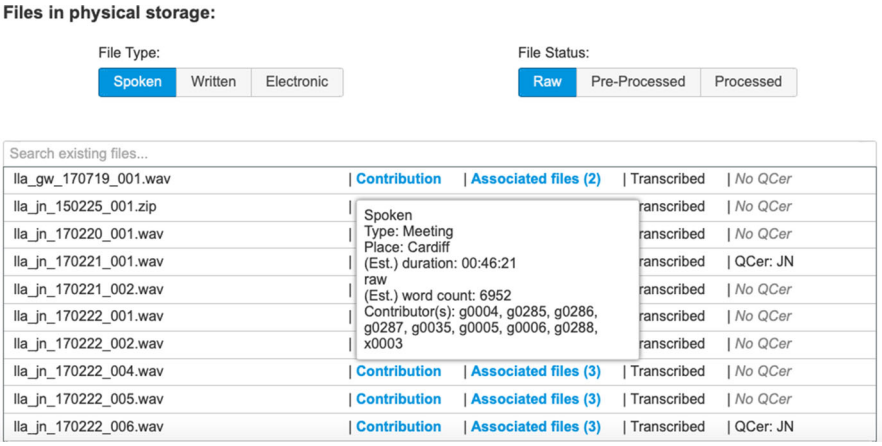


Fig. 3 A screenshot of the data management tool

encoding was correct. Further manual spot checks were made on up to 1000 words from 25% of all transcribed data files to ensure consistency of spelling and the use of apostrophes, etc. The resulting list of common misspellings was used to correct the entire sub-corpus automatically.

To support collaborative multi-user access by the team members (from researchers to transcribers) across different sites (including the multiple institutions involved in the project), an online data management tool was developed on top of the file storage system. It provides a graphical user interface that facilitates 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. Figure 3 illustrates typical user interaction with the data management tool.

3.5 Data processing

Once the data had been converted to plain text format, they were marked up automatically to add another layer of linguistic metadata that could be used to query the data. The CorCenCC team developed CyTag (Steven Neale et al. 2018), a suite of surface-level rule-based NLP tools for Welsh, based on the concept of 'constraint grammar' (Karlsson 1990; Karlsson et al. 1995) and implemented in Python. It supports text segmentation including sentence splitting and tokenization as well as part-of-speech (POS) tagging and lemmatization. It provides a bespoke solution for the basic linguistic preprocessing of Welsh including a tagset that is rich enough to capture idiosyncrasies of the language, including in its spoken form. The tagset contains a total of 145 fine-grained POS tags, which are mapped into 13 categories compliant with the EAGLES guidelines (Expert Advisory Group on Language Engineering Standards 1996). They include major syntactic categories ('noun', 'article', 'preposition', 'conjunction', 'numeral', 'adjective', 'adverb', 'verb',

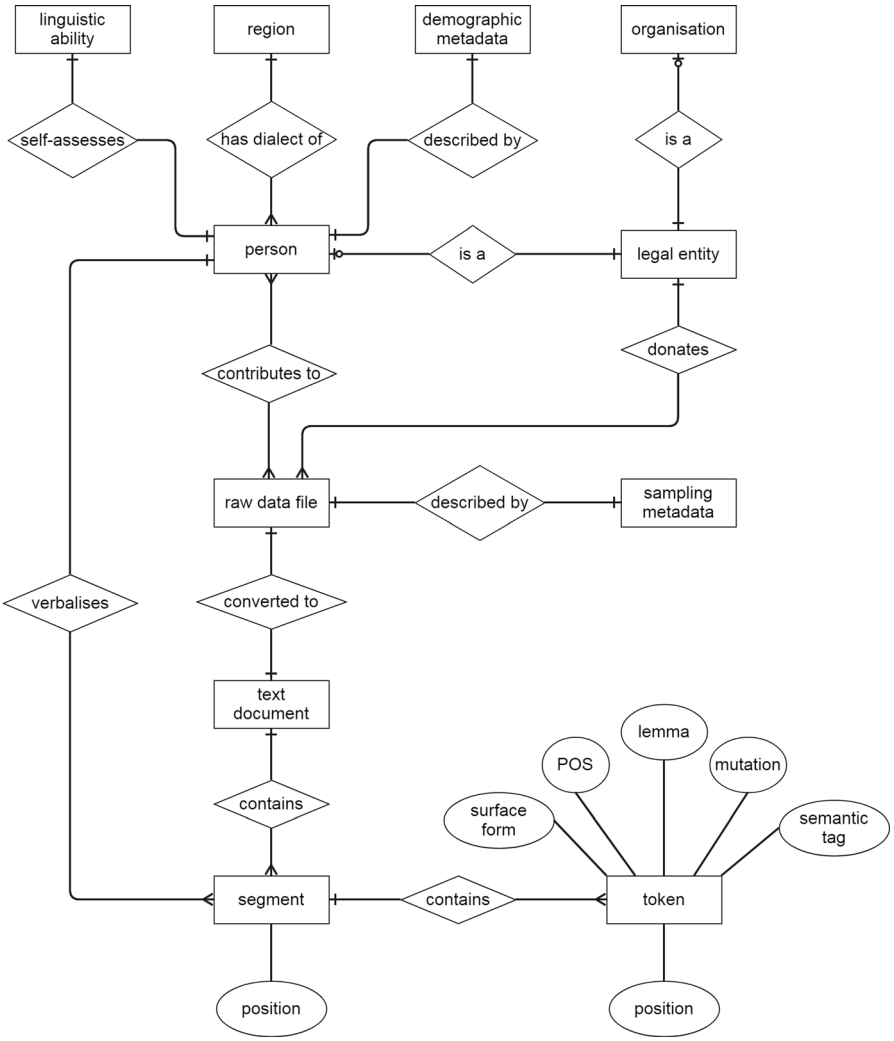


Fig. 4 Entity relationship model of the database used to store and manage the corpus

'pronoun', 'interjection', and 'punctuation') 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 fine-grained tags cover 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.

To facilitate the semantic analysis of Welsh language data on a large scale, all preprocessed data were further marked up by semantic categories. A semantic tagger, which was originally developed for English (Rayson et al. 2004), was successfully adapted for Welsh (Piao et al. 2018). CySemTagger combines a set of

lexical knowledge resources and disambiguation rules to map words and phrases into thesaurus type classifications of word senses. A total of 232 tags are arranged into 21 major categories including 'arts and crafts', 'science and technology', 'movement, location, travel and transport', 'numbers and measurement', etc.

All data processing outcomes were recorded by the data management tool. The processed data, together with the corresponding metadata, were now ready to be imported into a database designed to facilitate efficient retrieval of relevant data on demand.

3.6 Data storage

The corpus was stored and managed in a relational database where data could be accessed securely and concurrently by multiple users. The entity-relationship diagram shown in Fig. 4 illustrates the conceptual design of the database, in which, for the sake of simplicity, we do not show the attributes that correspond to the metadata fields described earlier in Sect. 3.2. The raw data file entity is central to the model. The corresponding table stores administrative metadata such as file name, its location, date of creation, permission of use, etc. The metadata related to the sampling framework is stored separately. The file itself is donated by a legal entity, which can be either an individual person or an organisation. One or more persons can make contributions to the content of the raw data file. For example, a journalistic interview will typically involve an interviewer and an interviewee, while a conversation recorded on the crowdsourcing app can involve multiple participants. For the data collected using the crowdsourcing app, the registered participants are anonymised, and described by metadata related to their demographics, linguistic ability and dialect.

Each raw data file was first converted to a plain text document, which was then processed by the Welsh NLP tools to segment the text into discrete units such as sentences and tokens. In the database, each token is stored separately with the corresponding attributes including surface form, POS tag, lemma, semantic tag and mutation (i.e. initial consonant mutation under specific morphological and syntactic conditions). All attributes are indexed individually for fast retrieval. Its position within a segment is also indexed. In combination with other attributes, this information can be used for the efficient computation of concordances, collocations and any other relevant patterns or statistics. Similarly, the segments themselves are ordered by their positions within a text document. When multiple persons are involved in a conversation represented by the document, each segment is attributed to the person who verbalised it originally.

3.7 Data access

To share the data online, we implemented a web-based interface to the database. The main reasons for creating a bespoke interface rather than re-using an existing solution such as CQPweb (Hardie 2012) were the requirements to tailor its functionality to the specific metadata of the CorCenCC corpus and its prospective users, and so it could be integrated with a bespoke pedagogic toolkit. To gather the

Table 5 User requirements survey

No.	Question
1	How do you describe your level of expertise in using corpus query tools?
2	What do you use corpus query tools for?
3	Which functionalities do you use most commonly?
4	What is your favourite corpus query tool(s) and why?
5	What functionalities would you like to see added to your favourite tool(s)?
6	What would improve the usability of your favourite tool(s)?

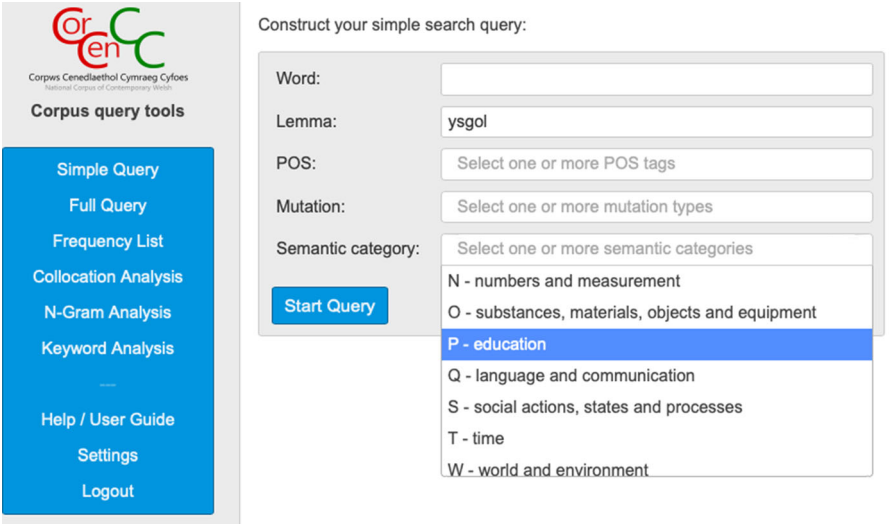


Fig. 5 The main menu

user requirements, we used social media to survey current users of corpora (see Table 5). A total of 62 individuals responded, and their input identified the key functionality requirements. In order of importance, they included: keywords in context (KWIC), concordancing, frequency lists, collocation analysis, visualisation, keyword analysis and *n*-gram analysis. We proceed by describing how these functionalities are supported by the CorCenCC interface, which features dual language controls in Welsh and English.

Upon landing on the home page, the registered users are allowed to log in. The registration is free and is required purely for monitoring the number of users engaging with the corpus. Once logged in, users are presented with a menu to help them navigate through various functionalities (see Fig. 5), as described in the remainder of this section.

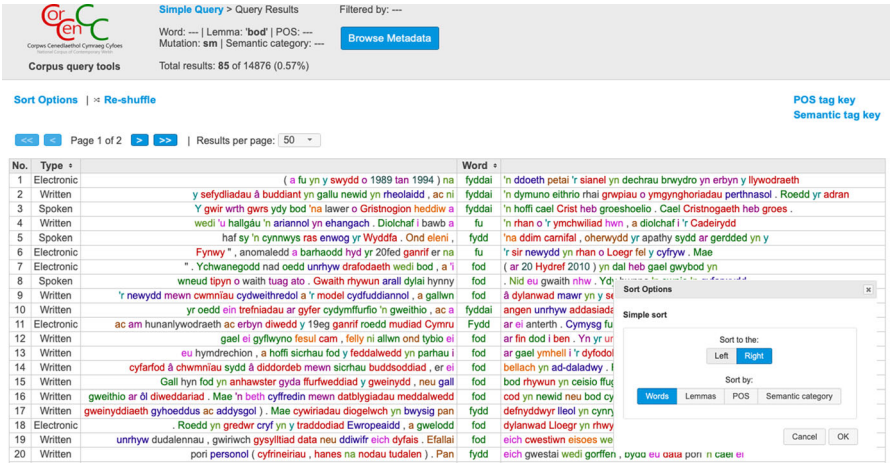


Fig. 6 The search results displayed as KWIC concordance lines

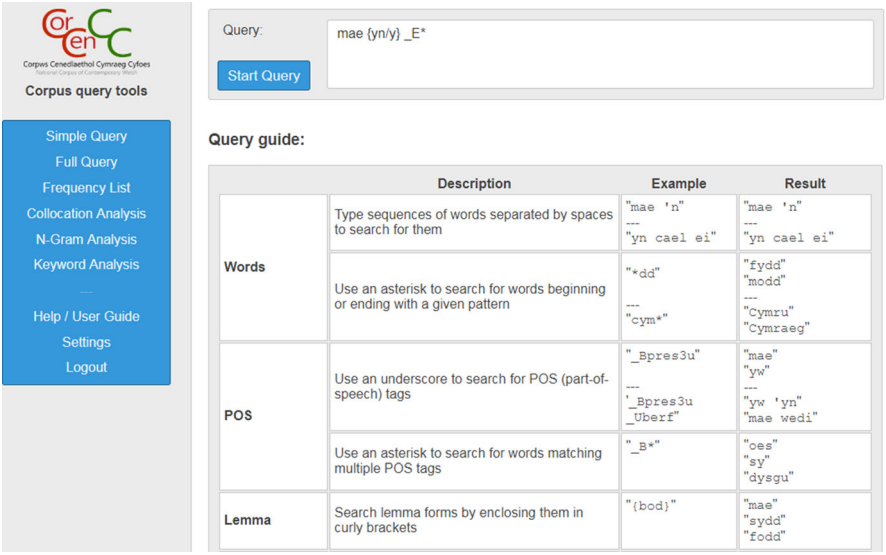


Fig. 7 A query designed to retrieve all occurrences of the word 'mae', followed by a word whose lemma is either 'yn' or 'y', followed by any noun

3.7.1 Simple query

Simple query provides a preformatted (or canned) query to allow novice users to familiarise themselves with the system and its search parameters. The user can search the corpus by providing a surface form of the word and/or its lemma. Optionally, the search can be further constrained by POS tags, mutation types and

semantic tags, which can be selected from a drop-down menu (see Fig. 5). The search results are displayed as KWIC concordance lines, where individual words are colour-coded according to their POS tags (see Fig. 6).

By default, the search results are ordered randomly. However, they can be sorted alphabetically or by language type (spoken, written and electronic), or re-shuffled. When sorting the results alphabetically, the user can choose the target to sort on, which can be a word to the left or to the right of the search term. They can then be sorted by their surface form, lemma, POS tag or semantic tag. In addition, the results can be filtered using the metadata, as described later.

3.7.2 Full query

Full query allows a user to formulate bespoke queries using a syntax that is based on the simple query syntax from CQPweb (Hardie 2012). Unlike simple queries, which focus on a single word at the time, the full query can be used to search for sequences of patterns (multi-word expressions) separated by spaces (see Fig. 7). As before, the search results are displayed as KWIC concordance lines.

3.7.3 Frequency list

The corpus data can be used to study the distribution of words and lemmas across different syntactic, semantic, discourse and demographic contexts. The frequency lists can be obtained for either words or their lemmas. The frequency calculations can be constrained by POS tags, mutation types[1] and/or semantic tags, in a manner compatible with that of a simple query described earlier (see Fig. 5). Figure 8 provides an example of the frequency analysis for soft mutated adjectives in the semantic category A: general and abstract terms (Termau cyffredinol a haniaethol).

3.7.4 Collocation analysis

The corpus data can be used to study the systematic co-occurrence of words (collocations) within a text window encompassing up to 7 words on either side of a given word (see Fig. 9). The affinity of collocates (i.e. cohesion) can be measured by a metric selected from a drop-down menu. Options include mutual information, MI3, Z-score and observed/expected, as defined in (Aston and Burnard 1998).

3.7.5 N-gram analysis

In addition to frequency lists, the corpus data can be used to study the distribution of n-grams ($2 \leq n \leq 7$), which can combine surface forms, their lemmas or POS tags. Figure 10 provides a frequency list for all 4-grams based on the surface forms.

[1] In mutation, certain consonants change in predictable ways according to the phonological or morphological environment. Welsh has three mutations that apply to the starts of (some) words.

List of: **words**

Constrained to: POS (**Egu**) | Mutations (**sm**) | Semantic category (**A**)

Total results: **39** unique word types (From 14876 tokens)

No.	Word	POS	Count	Frequency
1	gilydd	Egu	8	0.05%
2	ben	Egu	5	0.03%
3	ddewis	Egu	5	0.03%
4	fywyd	Egu	4	0.03%
5	ganlyniad	Egu	3	0.02%
6	gysylltiad	Egu	3	0.02%
7	Lafur	Egu	3	0.02%
8	beth	Egu	2	0.01%
9	ddylanwad	Egu	2	0.01%
10	derfyn	Egu	2	0.01%
11	dro	Egu	2	0.01%
12	wall	Egu	2	0.01%
13	becyn	Egu	1	0.01%
14	bennawd	Egu	1	0.01%
15	ddiffyg	Egu	1	0.01%
16	ddioddef	Egu	1	0.01%
17	ddirywiad	Egu	1	0.01%
18	ddrych	Egu	1	0.01%
19	doriad	Egu	1	0.01%
20	drawsnewidiad	Egu	1	0.01%

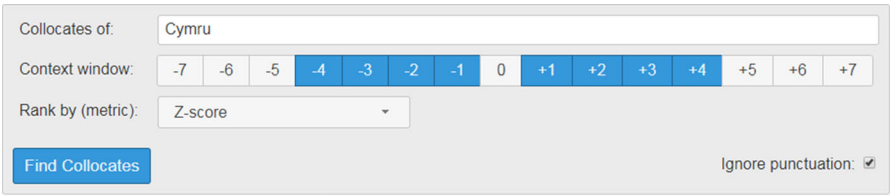
Fig. 8 A sortable frequency list

3.7.6 Keyword analysis

The corpus data can be used to study the distribution of words across different subsets of the data, as defined by selecting different values for the corpus metadata. For example, Fig. 11 illustrates how the user can compare the use of spoken language at different places in Wales. Such a query can highlight the differences in usage frequency according to the log-likelihood measure and a level of statistical significance.

3.7.7 Metadata selection

Metadata can be selected from a dialogue box with three tabs, for the spoken, written and electronic language respectively. Each tab can then be used to further constrain the data based on metadata fields that are appropriate for the corresponding type of language (see Fig. 12).



Collocates of: **Cymru**

Context window: **-4 to +4**

Ranked by (metric): Z-score

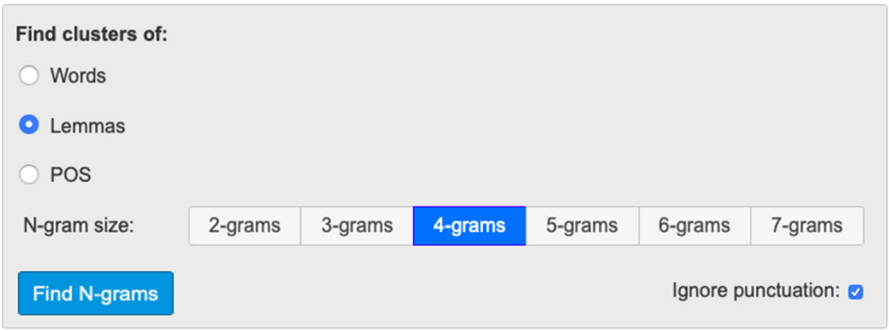
Total results: **338** cydleoliadau (From 87 instances of 'Cymru')

No.	Collocate	Count	Freq. (with 'Cymru')	Freq. (total)	Rank
1	A	23	26.44%	0.15%	12.83
2	Llywodraeth	11	12.64%	0.07%	12.40
3	mae	13	14.94%	0.09%	12.38
4	gan	10	11.49%	0.07%	10.76
5	'r	18	20.69%	0.12%	7.90
6	yw	7	8.05%	0.05%	6.93
7	ar	12	13.79%	0.08%	6.34
8	i	17	19.54%	0.11%	6.25
9	Cynulliad	7	8.05%	0.05%	5.42
10	Hefyd	5	5.75%	0.03%	4.11
11	wedi	6	6.90%	0.04%	3.92
12	BBC	5	5.75%	0.03%	3.33
13	cenedlaethol	4	4.60%	0.03%	2.94
14	ond	4	4.60%	0.03%	2.03
15	S4C	3	3.45%	0.02%	1.98
16	tywysog	4	4.60%	0.03%	1.98
17	Roedd	3	3.45%	0.02%	1.91
18	bobl	3	3.45%	0.02%	1.90
19	ac	7	8.05%	0.05%	1.84
20	Bu	3	3.45%	0.02%	1.83

Fig. 9 Collocation analysis

4 Results

An important consideration for the development of the corpus was checking that it is fit for purpose. We evaluated the usability and functionality of the web-based interface amongst linguists, who comprise our target user group. Three domain experts agreed to carry out road testing. They were linguists at three different higher education institutions, and each more than 15 years of experience in corpus linguistics. None of the participants had used our web-based platform previously and none were speakers of Welsh. Separate procedures, not described here, are being used to evaluate the usability of the content of CorCenCC itself and this assessment is focusing primarily on Welsh speaking users. Ethical approval was obtained using Cardiff University's standard procedures, and the participants gave informed consent. They were instructed to open a link to our online tool and to explore its functionality. There were no predetermined tasks or goals to be met. However, the participants were asked to be exhaustive in utilising all available features. The think-aloud protocol (Jääskeläinen 2010) was followed to capture rich



4-grams of type: **Lemmas**

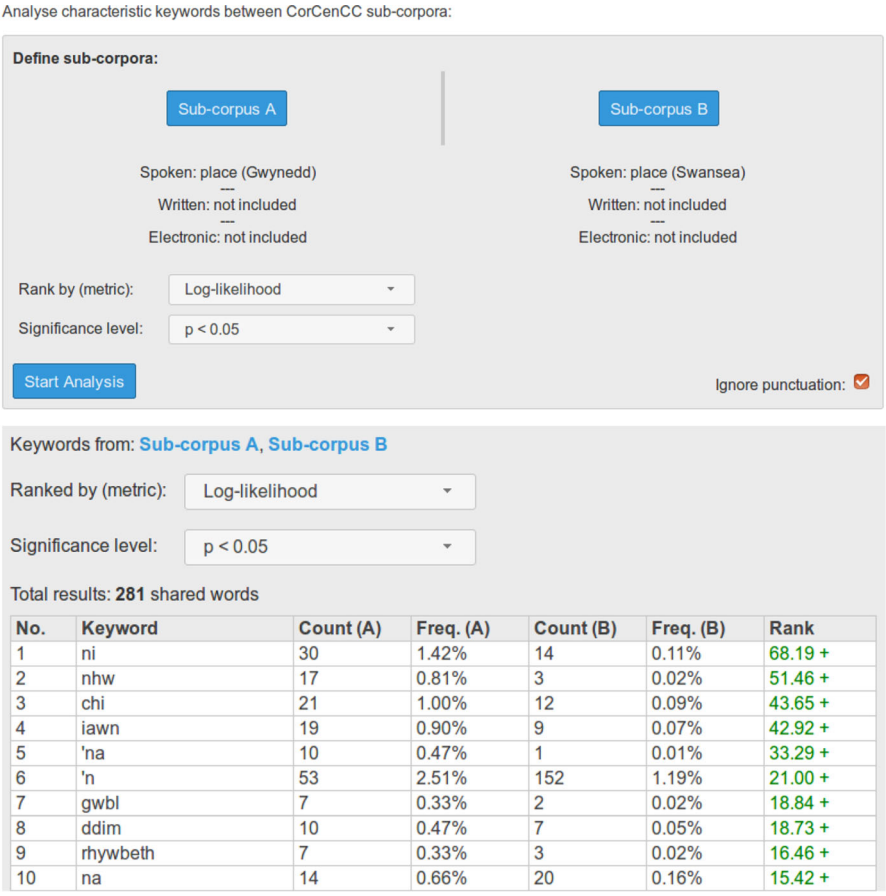
Total results: **8759** unique 4-grams

No.	4-gram	Count	Frequency
1	bod yn ymwneud â	8	0.09%
2	bod yn cael ei	7	0.08%
3	bod yn cael eu	7	0.08%
4	bod y rhaglen yn	4	0.05%
5	dim carnifal yn Llanberis	4	0.05%
6	llyfr bod yn ymwneud	4	0.05%
7	ar hwn o pryd	3	0.03%
8	bod ar cael i	3	0.03%
9	bod yn rhaid i	3	0.03%
10	bod yn rhan o	3	0.03%
11	bod yn tebyg y	3	0.03%
12	bod yn ymddiried yn	3	0.03%
13	dwyrain Ewrop a y	3	0.03%
14	enillydd Awstralaidd un erioed	3	0.03%
15	hanes bro a teulu	3	0.03%
16	olrhain hanes bro a	3	0.03%
17	os na bod wedi	3	0.03%
18	yn cael ei defnyddio	3	0.03%
19	yn mawr na hwnnw	3	0.03%
20	yn siŵr eich bod	3	0.03%

Fig. 10 *N*-gram analysis

qualitative data. Each user test took approximately 1 h during which we recorded a participant's verbal feedback together with their activities on the screen. In addition, we interviewed the participants post-task. We used the information to conduct a thematic analysis (Braun and Clarke 2012). No predefined themes were used. Instead, an inductive approach was used to produce the codes (see Table 6). These codes helped us to capture and characterise the underlying issues in a constructive way, so they could be addressed as generically as possible.

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



No.	Keyword	Count (A)	Freq. (A)	Count (B)	Freq. (B)	Rank
1	ni	30	1.42%	14	0.11%	68.19 +
2	nhw	17	0.81%	3	0.02%	51.46 +
3	chi	21	1.00%	12	0.09%	43.65 +
4	iawn	19	0.90%	9	0.07%	42.92 +
5	'na	10	0.47%	1	0.01%	33.29 +
6	'n	53	2.51%	152	1.19%	21.00 +
7	gwbl	7	0.33%	2	0.02%	18.84 +
8	ddim	10	0.47%	7	0.05%	18.73 +
9	rhywbeth	7	0.33%	3	0.02%	16.46 +
10	na	14	0.66%	20	0.16%	15.42 +

Fig. 11 Keyword analysis

participants agreed that they were likely to adopt the system and recommend it to other linguists. Nonetheless, the participants identified specific functionality that they would like to see improved, either by addressing system errors or through producing new functionalities and/or adding features to meet additional requirements producing new functional requirements. Examples included Chi squared testing, translating features between English and Welsh and saving the search results (e.g. concordances) locally.

All participants faced one key issue with a non-functional aspect of the system. Operations such as searching incurred a relatively long time delay between the user's input and the system's output. Under uncertainty about the cause of a delay, the users opted to terminate the process rather than waiting any longer. From this, we learned that keeping the response time to a minimum would be a crucial part of successful interaction with, and adoption of, the system. It was concluded that, to

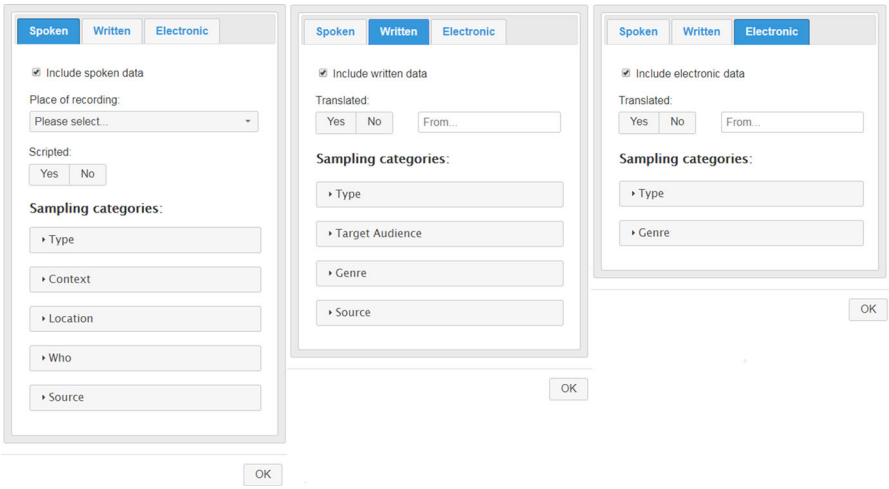


Fig. 12 Metadata selection

accommodate likely lags, better status feedback should be provided to the users to keep them engaged.

Our participants commented on their uncertainty about how the system operated internally and how the outputs were derived in some cases. A prime example was that they would have liked more information about the ranking score, in terms of its calculation and relevance to the user (see Figs. 9 and 11). The users also reported uncertainty regarding some of the descriptors and outputs from the system. Specific examples included not knowing what nasal mutations were, and being confused by some of the POS tags (see Fig. 7). Suggestions such as mouseover explanations were given as a way of enhancing the users' understanding. In response to the usability testing, we addressed the development bugs that produced error messages. Other codes provided in Table 6 will be used to make improvements in the next version of the system.

Road-testing with Welsh-speakers is expected to identify other types of issue, related to the accuracy of the language mark-up. While Welsh-speakers will fare better than our system road-test participants with regard to recognising features relevant to the analysis of Welsh, they may have less knowledge of, and confidence in, using the interface. Our staged approach to road-testing, with different types of user, has been specifically aimed to address different types of challenge discretely. For example, the sorts of basic glitches identified by the corpus experts needed to be fixed before the corpus was trialled by people less able to articulate such difficulties.

Table 6 Thematic analysis

Theme	Code
Technical functional	Error message received
	Language switching on logging in
	Concordances cannot be saved
	Translate feature needed
	POS tag showing in parallel to other information
	Chi squared functionality not working
	logDice calculation not provided
	n-gram selection with words
	Context window not given (i.e. revealing extending passages of the concordance lines)
Technical non-functional	Slow response
Black box functionality	Sorting confusion
	No results provided
	Ranking calculation unclear
Uncertainty of descriptors	Search capabilities
	POS
	Nasal mutations specific to Welsh
	Frequency
	Count
Usability	Selection difficulty
	Sign-up difficulty
	Logo clicks

5 Conclusions

The construction of a new corpus infrastructure is a major undertaking. Where many researchers are able to gather new text and then analyse it using existing software, our task was a very different one. Specifically, we were simultaneously addressing several major challenges important for progressing corpus linguistic research. Firstly, we have created the largest and most comprehensive corpus of the Welsh language, reflective of language use across communication types, genres, speakers, language varieties (regional and social) and contexts. This new resource will revolutionise the understanding of a language that has not only a long cultural heritage but also a striking recent revival history that has shaken up traditional approaches to, and perceptions of, standardisation.

Secondly, we have necessarily designed from scratch much of the underlying computational infrastructure for tagging and analysing the language, given that Welsh has many features (including significant regional and register variation) that do not transfer easily from English. 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 with limited Welsh will be encouraged by this means to buy wholeheartedly into the language as not only a source of information but the medium through which it can be studied. At the same time, the availability of an English language interface as well 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 as a second or foreign language worldwide.

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. In this way, we both broaden the potential utility and impact of our own work and act as a catalyst for work on many other languages.

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. To give one example, the crowdsourcing app enables ‘live’ user-generated spoken data collection data collection using any Internet-enabled device, including both mobile devices and an interactive website. The crowdsourcing application 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 necessary manual tagging, even for a language with a relatively small cohort of fluent speakers.

Given the restricted timespan and scale of the initial CorCenCC project (3.5 years, with a pre-specified funding pot), the focus was creating an initial dataset of 10 M words of written and spoken contemporary Welsh language. Therefore, our intention is to extend the size of the corpus as soon as possible. With the infrastructure already in place, growing the corpus is not a major challenge, provided further funding can be secured.

Authors' contributions DK, SN, LA and IS co-designed the computational infrastructure for the CorCenCC corpus. SN implemented the system. FL conducted the usability study. DK and IS drafted the manuscript. All authors read and approved the final manuscript.

Funding This work has been funded by the UK Economic and Social Research Council (ESRC) and Arts and Humanities Research Council (AHRC) as part of 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).

Availability of data and materials The CorCenCC corpus can be accessed at http://www.corcencc.org/. The data are available under CC-BY-SA v4 license.

Code availability The code from the CorCenCC project is shared at http://www.corcencc.org and https://github.com/CorCenCC under GPL v3 license.

Compliance with ethical standards

Conflicts of interest All authors declares that they have no conflict of interest.

Ethics approval Ethical approval was gained from Andrew Edgar, Ethics Officer in the School of English, Communication and Philosophy, Cardiff University.

References

Adolphs, S., Knight, D., Smith, C., & Price, D. (2020). Crowdsourcing formulaic phrases: towards a new type of spoken corpus. *Corpora*, *15*(1), in press.

Anthony, L. (2014). AntConc (Version 3.4.3). Waseda University. https://www.laurenceanthony.net/software/antconc/. Accessed 27 July 2020.

Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Diaz de Ilarraza, A., et al. (2007). ZT corpus: Annotation and tools for Basque corpora. Paper presented at the Corpus Linguistics Conference, Birmingham.

Aston, G., & Burnard, L. (1998). *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*(1), 1–16.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243–257.

Boleda, G., Bott, S., Villanueva Meza, R. M., Castillo, C., Badia, T., & López, V. (2006). CUCWeb: A Catalan corpus built from the web. Paper presented at the the 2nd International Workshop on Web as Corpus, Trento.

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving: an introduction and cases. *Convergence, 14*(1), 75–90. https://doi.org/10.1177/1354856507084420.

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. Camic, D. Long, A. Panter, D. Rindskopf, & K. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004.

Carter, R., & McCarthy, M. (1997). *Exploring Spoken English*. Cambridge: Cambridge University Press.

Carter, R., & McCarthy, M. (2004). Talking, creating: interactional language, creativity, and context. *Applied Linguistics*, 25(1), 62–88. https://academic.oup.com/applij/article/25/1/62/149094.

Davies, M. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25(4), 447–464. https://academic.oup.com/dsh/article/25/4/447/997323.

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide, 36*(1), 1–28. https://doi.org/10.1075/eww.36.1.01dav.

Deuchar, M., Davies, P., Herring, J., Parafita Couto, M., & Carter, D. (2014). Building bilingual corpora. In E. M. Thomas & I. Mennen (Eds.), *Advances in the Study of Bilingualism* (pp. 93–111). Bristol: Multilingual Matters.

Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata principles and practicalities. *D-lib Magazine, 8*(4), 1082–9873.

Ellis, N. C., O'Dochartaigh, C., Hicks, W., Morgan, M., & Laporte, N. (2001). Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh. https://www.bangor.ac.uk/canolfanbedwyr/ceg.php.en. Accessed 27 July 2020.

ESRC Centre for Research on Bilingualism (2020). Bangor Siarad. http://bangortalk.org.uk/. Accessed 27 July 2020.

Estellés-Arolas, E., & González-Ladrón-De-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science, 38*(2), 189–200. https://doi.org/10.1177/0165551512437638.

Expert Advisory Group on Language Engineering Standards (1996). EAGLES guidelines. http://www.ilc.cnr.it/EAGLES/browse.html. Accessed 27 July 2020.

Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380–409. https://www.ingentaconnect.com/content/jbp/ijcl/2012/00000017/00000003/art00004.

Jääskeläinen, R. (2010). Think-aloud protocol. In Y. Gambier, & L. van Doorslaer (Eds.), *Benjamins Handbook of Translation Studies, Volume 1* (pp. 371-373). Amsterdam/Philadelphia John Benjamins.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). *'The TenTen Corpus Family' the 7th International Corpus Linguistics Conference* (pp. 125–127). UK: Lancaster.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. Paper presented at the the 13th International Conference on Computational Linguistics (COLING), Helsinki.

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

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., et al. (2014). The Sketch Engine: ten years on. *Lexicography, 1,* 7–36. https://doi.org/10.1007/s40607-014-0009-9.

Knight, D. (2011). The future of corpus linguistics. *Brazilian Journal of Applied Linguistics, 11*(2), 391–416.

Knight, D., Adolphs, S., & Carter, R. (2013). Formality in digital discourse: a study of hedging in CANELC. In J. Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics* (pp. 131–152). Netherlands: Springer. http://orca.cf.ac.uk/78844/.

Knight, D., Fitzpatrick, T., & Morris, S. (2017). CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes—The National Corpus of Contemporary Welsh): An overview. (Paper presented at the the Annual British Association for Applied Linguistics (BAAL) Conference, Leeds, UK).

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, B., Anthony, L., Cobb, T., Deuchar, M., Donnelly, K., McCarthy, M., & Scannell, K. (2020). CorCenCC: (Corpws Cenedlaethol Cymraeg Cyfoes – The National Corpus of Contemporary Welsh): A community driven approach to linguistic corpus construction. https://urldefense.proofpoint.com/v2/url?u=http-3A__www.corcencc.org_explore&d=DwIGaQ&c=vh6FgFnduejNhPPD0fl_yRaSfZy8CWbWnIf4XJhSqx8&r=r2aSgYn6PHMQXXmeBiKsnvfFG9T9U5fmdQ67xEVmgo0&m=vENIMNwG0whbhOk5BSn93DDjUNbEHZkw7lOkhXD_WU&s=XozS5TJ9oHkKopOV4ytseYQ1EtUmmW8_QjYuqwg1hcg&e=. Accessed 27 July 2020.

Kucera, K. (2002). The Czech National Corpus: principles, design, and results. *Literary and Linguistic Computing, 17*(2), 245–257.

Kupietz, M., Lüngen, H., Kamocki, P., & Witt, A. (2018) 'The German reference corpus DeReKo: New developments—new opportunities' *the Eleventh International Conference on Language Resources and Evaluation* (pp. 4354–4360). Miyazaki, Japan. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/7491.

Leech, G. (2014). The state of the art in corpus linguistics. In K. Aijmer, & B. Altenberg (Eds.), *English Corpus Linguistics* (pp. 20–41). Routledge.

Love, R. (2020). *Overcoming challenges in corpus construction: The spoken British National Corpus 2014*. Abingdon: Routledge.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Mahwah: Lawrence Erlbaum Associates.

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. https://benjamins.com/catalog/ijcl.22.3.01mce.

Meyer, C. F. (2002). *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Mozilla (2020). Common Voice. https://voice.mozilla.org/. Accessed 27 July 2020.

Neale, S., Donnelly, K., Watkins, G., & Knight, D. (2018). Leveraging lexical resources and constraint grammar for rule-based part-of-speech tagging in Welsh. Paper presented at the the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki.

Neale, S., Spasić, I., Needs, J., Watkins, G., Morris, S., Fitzpatrick, T., et al. (2017). The CorCenCC crowdsourcing app: A bespoke tool for the user-driven creation of the national corpus of contemporary Welsh. Paper presented at the Corpus Linguistics Conference, Birmingham.

Office for National Statistics (2011). UK census. https://www.ons.gov.uk/census/2011census. Accessed.

Piao, S., Rayson, P., Knight, D., & Watkins, G. (2018). Towards a Welsh semantic annotation system. Paper presented at the the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki.

Prys, D., & Jones, D. B. (2018). Gathering data for speech technology in the welsh language: A case study. (aper presented at the the LREC Workshop on Collaboration and Computing for Under-Resourced Languages Sustaining knowledge diversity in the digital age, Miyazaki.

Rayson, P., Archer, D., Piao, S., & McEneryb, T. (2004). The UCREL semantic analysis system. Paper presented at the the Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP tasks at the 4th International Conference on Language Resources and Evaluation (LREC), Lisbon.

Rees, M., Watkins, G., Needs, J., Morris, S., & Knight, D. (2017). Creating a bespoke corpus sampling frame for a minoritised language: CorCenCC, the National Corpus of Contemporary Welsh. Paper presented at the the Corpus Linguistics Conference, Birmingham.

Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In C. Fairon, H. Naets, A. Kilgariff, & G.-M. De Schryver (Eds.), *Building and Exploring Web Corpora, Proceedings of the 3rd Web as Corpus Workshop* (p. 182). Presses universitaires de Louvain. http://crubadan.org/languages/cy.

Schmidt, T. (2014). The database for spoken German—DGD2. Paper presented at the the 9th International Conference Language Resources and Evaluation, Reykjavik, Iceland.
Simpson-Vlach, R. C., & Leicher, S. (2006). The MICASE handbook: A resource for users of the Michigan corpus of academic spoken English. University of Michigan Press.
Sinclair, J. (2005). Corpus and text—basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxford: Oxbow Books.
Tadić, M. (2002). Building the Croatioan National Corpus. Paper presented at the The third International Conference on Language Resources and Evaluation (LREC), Las palmas.
Weinberger, S. H. (2020). The speech accent archive. http://accent.gmu.edu/. Accessed 27 July 2020.
Williams, B. (1999). A Welsh speech database: Preliminary results. Paper presented at the the 6th European Conference on Speech Communication and Technology (EUROSPEECH), Budapest.