Creating a Bespoke Corpus Sampling Frame for a Minoritised Language:
CorCenCC, the National Corpus of Contemporary Welsh

Mair Rees (Swansea University, UK), Gareth Watkins (Cardiff University, UK), Jennifer Needs (Swansea University, UK), Steve Morris (Swansea University, UK) and Dawn Knight (Cardiff University, UK)

Overview

In this paper, we discuss the steps taken to create a bespoke sampling frame to use when constructing a national corpus for a minoritised language, in this instance Welsh. We illustrate the processes involved in designing this sampling frame, with emphasis on how decisions were reached to reflect this context and how they are different to those which might be made in the construction of a sampling frame for a majority language such as English. The processes adhered to, and the decisions made as part of this design process, are potentially of value to other corpus linguists who may be interested in creating corpora for other minoritised languages.

As is widely documented, the design and construction principles used to build corpora are often locally determined (Conrad, 2002: 77), depending on the vision for the corpus and the questions that it is intended to help us answer. CorCenCC[i] (Corpws Cenedlaethol Cymraeg Cyfoes – The National Corpus of Contemporary Welsh) is an interdisciplinary, collaborative project, whose vision is to construct the first general principled corpus of contemporary Welsh.

CorCenCC will contain 10 million words by the end of the project, comprising 4 million each from spoken and written sources and 2 million from digitally mediated sources (e-language). Data will be drawn from a variety of contexts, ranging from formal (e.g. political documents, televised interviews and formal letters) to less formal ones (e.g. diaries, phone calls and text messages). A provisional outline for the distribution of this data is shown in Figure 1.

This paper will examine some key issues, questions and considerations involved in creating a bespoke sampling frame for the Welsh language.

Spoken language

The strategy used to collect spoken data for CorCenCC is adapted from CANCODE's 'genre' approach, which 'tries to seek a balance between speaker, environment, context and recurrent features' (McCarthy, 1998: 8), in combination with the BNC's 'context-governed' model (Crowdy, 1993: 259). CANCODE includes data from five contexts – Transactional, Professional, Pedagogical, Socialising and Intimate ('Private' in CorCenCC's terms) – but omits contexts associated with more formal language due to its focus on 'unrehearsed, non-formal talk' (McCarthy, 1998: 9). As a general corpus, CorCenCC does not have the same focus as CANCODE. It is important, therefore that both formal and informal language are represented within the corpus, particularly given that the syntax and morphology of Welsh vary considerably according to degrees of formality. For this reason, two contexts from the BNC's model – Public/Institutional and Leisure ('Media' in CorCenCC's terms) – have been added to

CorCenCC's spoken sampling frame, allowing it to reflect more fully the complexity of the Welsh language.

Spoken

Cyd-destun / *Context*	%	Geiriau / *Words*
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**

Written

Cyfrwng / *Medium*	%	Geiriau / *Words*
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**

E-Language

Math / *Type*	%	Geiriau / *Words*
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**

Figure 1: Provisional sampling frame for CorCenCC

Sampling from the seven contexts shown in Figure 1 will enable CorCenCC to represent all the different genres that are spoken and heard, including the language of the media, public and institutional language, the language of the workplace and of education, as well as personal and social use of Welsh. Data will also be collected from speakers 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 Wales' diversity not only of genres but also of Welsh speakers themselves.

Written language

CorCenCC's written sub-corpus is constructed on a genre-based framework in order to produce an 'information-rich, user-friendly resource' as outlined in Lee (2001: 63). The key challenges (at sociolinguistic, socio-political and practical levels) in compiling a sampling frame for written language, a sampling frame which is both balanced and representative of the Welsh language community, are illustrated in this paper. Firstly, there is a relatively disproportionate number of children's books which are published each year, for both demographic and commercial reasons (Rosser, 2012: 1). Secondly, the complexities of accurately representing the range and distribution of literature for

adult and child second-language learners at various levels of expertise. Thirdly, are the broader issues around readership, authorship, demography and dialect which again have impact on the balance and representativeness of the corpus.

These challenges were considered when designing the sampling frame for the written sub-corpus. For example, the medium 'Books' includes books written for adult learners, children, and child learners. It is anticipated that this will be of assistance to pedagogical end-users in the future, as these sub-sets will demonstrate the forms of Welsh which learners and children are most likely to come across when reading. Place of publication will be recorded in the metadata and a broad sample of written material will be sought overall to ensure that no one publisher dominates. The medium 'Magazines' includes the genre 'papurau bro', comprising of 60 local community Welsh-language newspapers/newsletters covering most areas of Wales which will capture a range of dialects and levels of proficiency.

E-language

It is only fairly recently that General, 'National' corpora have sought to include e-language data. According to Knight et al. (2014: 30) 'while it is widely acknowledged that we live and communicate in a ubiquitous, digital world, the ways in which we actually do this, across multiple resources, remains an underexplored area of research in corpus linguistics as there is a lack of appropriate resources in existence to enable us to do this'.

CorCenCC will reflect modern usage of language and the use of electronic mediums to facilitate communication in the Welsh language context.

However, while it appears that e-language is all pervasive in English, this is not the case for Welsh. For example, there was a provisional plan to include language data from discussion boards in CorCenCC, but only four boards and forums conducted through the medium of Welsh were identified. Of those four, three had been dormant for many years. Clearly, care needs to be taken when deciding which elements should be included. Including only one community of people to represent an e-language type appears ill advised, thus discussion boards will not be included in the CorCenCC corpus.

Research by Beaufort Research (2013: 28) suggests that there is further disparity between the number of e-mails and SMS messages communicated in Welsh compared to English, even among Welsh speakers. 42% of Welsh speaking respondents to the Beaufort Research survey had sent a Welsh language e-mail in the previous month, with 72% having sent at least one in English. Furthermore, only 44% of respondents had sent a Welsh language SMS in the previous month. While this supports the inclusion of these e-language types in the corpus, it also suggests that the number of allocated words for these e-language types should be proportionally less than had the sampling frame been designed for an English corpus.

Each e-language type is categorised in respect of its general topic, purpose or content. Blogs and websites will be split into six categories, each containing 100,000 words, as illustrated in Figure 2.

Code	Topic
A	News, Media and Current Affairs, Politics Business and Finance Weather and the Environment Online Shopping
B	Religion Language Culture, Literature and the Arts Teaching, Academia and Education
C	Technology, Computers and Gaming Fashion and Beauty Hobbies and Pastimes Travel Cookery
D	Music Sport Gigs and Events
E	Celebrity news and gossip TV and Film Humour
F	Parenting and Family Life Health and Wellbeing Personal and Daily Life

Figure 2: CorCenCC blog and website topics.

This thematic framework is broadly based on the one devised by the team constructing CANELC (Cambridge and Nottingham eLanguage Corpus) (see Knight et al. 2014), but with some important modifications to ensure that it is appropriate for the Welsh language context. For instance, religion was not included as a category in the CANELC framework. However, brief research conducted at the start of the CorCenCC project identified that a number of blogs were dedicated to discussing religious topics. This suggests that religion is an important domain in Welsh, and that it should therefore be included in the CorCenCC thematic framework. Of course, this framework is not appropriate for more private e-language types such as SMS and e-mail, which have been split according to the context in which these messages were sent, namely either business or personal.

In this presentation, the peculiarities of e-language use in the Welsh language context are explored, in respect of type and topic, both of which will be discussed in more detail.

References

Beaufort Research. (2013). *Ymchwilio i ddefnydd iaith siaradwyr Cymraeg yn eu bywyd pob dydd [Exploring Welsh speakers' language use in their daily lives]*. Retrieved from http://gov.wales/docs/dcells/research/130808-wels-lang-research-cy.pdf

Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics,* 22, 75-95.

Crowdy, S. (1993). Spoken Corpus Design. *Literary and Linguistic Computing* 8(4), 259-265.

Knight, D., Adolphs, S. & Carter, R. (2014). CANELC – constructing an e-language corpus. *Corpora Journal* , 9(1), 29-56.

Lee, D. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology,* 5(3), 37-72.

McCarthy, M. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

Rosser, S. M. (2012). Language, culture and identity in Welsh children's literature: O. M. Edwards and Cymru'r Plant 1892-1920. in Nic Congáil, R. (ed.), *Codladh Céad Bliain: Cnuasach Aistí ar Litríocht na nÓg* (pp. 223-251). Dublin: Leabhair COMHAR, pp. 223-251.

Welsh Books Council. (2012). *Buying and Reading Welsh-language Books: Welsh Speakers Omnibus Survey 2012 – Report of survey findings*. Retrieved from http://www.cllc.org.uk/ni-us/cyhoeddiadau-publications/ymchwil-research

---

[i] CorCenCC is funded by the Economic and Social Research Council (ESRC) and the Arts and Humanities Research Council (AHRC). Ref ES/M011348/1.