"How will you make sure the material is suitable for children?": User-informed design of Welsh corpus-based learning/teaching tools

Jennifer Needs (Swansea University, UK), Dawn Knight (Cardiff University, UK), Steve Morris (Swansea University, UK), Tess Fitzpatrick (Swansea University, UK), Enlli Thomas (Bangor University, UK) and Steven Neale (Cardiff University, UK)

The CorCenCC project (Corpws Cenedlaethol Cymraeg Cyfoes – The National Corpus of Contemporary Welsh) is breaking new ground, in that it is creating the first ever large-scale corpus of Welsh and using pioneering community-driven methods, and also in that pedagogical corpus tools are part of CorCenCC's design from the very outset. The 10-million-word corpus will be a unique resource for Welsh language translators, lexicographers, publishers, policy-makers, language technology developers, researchers, and for those learning and teaching Welsh. This paper looks at the needs of this last group of end-users, and the challenges and opportunities arising from the task of developing a bespoke online pedagogical toolkit which works directly with the corpus data.

Welsh is one of the more privileged of the world's lesser-used languages, in that it has government recognition and support, a rich literary tradition, dedicated television and radio stations, and is an important part of the education system in Wales. The Welsh language is taught as a second language in primary and secondary schools and in post-16 education, and it is also possible to attend Welsh-medium education from nursery right through to university-level. However, with no comprehensive corpus of Welsh, pedagogical materials have been based largely on intuition in terms of the grammar and vocabulary items to target, and authentic listening and reading passages have been few and far between.

CorCenCC's design is user-informed, and consultation with Welsh teachers and tutors at all stages of education has shown how beneficial the corpus will be as a bank of authentic material. CorCenCC will include spoken, written and electronic language from all contexts where the language is used, e.g. in private, for socialising, for business and work, in education, in the media, and in public. It will include examples of news headlines, personal and professional letters and emails, academic writing, formal and informal speech, and even text messaging. The Welsh language varies considerably depending on formality, so CorCenCC will be a valuable resource for those teaching and learning about different genres and registers in Welsh.

CorCenCC will be freely available to the general public online via a generic user interface, but the same website will also include a second interface designed specifically for educational purposes. Both interfaces will work with the entire corpus data, but the educational interface is tailored for use by teachers and tutors from primary school to HE and adult education for use in their Welsh classes, and also at the pupils/students/learners themselves from GCSE-level up. During initial consultations with primary school teachers about CorCenCC's potential role in their lessons, a concern arose which represents a tension between the material's authenticity on the one hand, and considerations of appropriate classroom language on the other: "How will you make sure the material is suitable for children?". This paper addresses the impact of this concern on the development of CorCenCC's pedagogical interface.

Profanity

Perhaps the most obviously 'unsuitable' material for children is that which includes profane or offensive language. Because the corpus data will be uncensored, if teachers are to use CorCenCC as a resource in their classrooms, the special pedagogical interface must have an in-built tool to filter out corpus examples that contain such language. Teacher feedback thus far has pointed towards an online resource which might be used as a starting point for identification of unsuitable language, but it is by no means comprehensive and, given the bilingual situation in which the Welsh language exists, any proposed 'filter' must be capable of dealing with unsuitable vocabulary in English as well as Welsh. This may not be as simple as listing vocabulary items and their translations! Not all terms which are offensive in one language are necessarily offensive in translation. There is also the additional challenge of consonant 'mutation' in Welsh, whereby a word could potentially be spelt four different ways, depending on context – e.g. *clinig* (clinic), *dy glinig* (your clinic), *fy nghlinig* (my clinic), *ei chlinig* (her clinic).

Subject matter

Clinig is given as an example here, as it is evidently not an offensive or unsuitable word in itself. However, since CorCenCC will include spoken, written and e-language from a wide range of genres/subjects, it is likely that there will be coverage of certain topics which might be deemed unsuitable for children – discussion of personal medical conditions, for example, or perhaps discussion of alcohol consumption or sex. In addition to a 'profanity filter', CorCenCC will need a strategy for identifying and marking up 'adult' subject matter, so that corpus search results obtained via the pedagogical interface can be tailored to exclude passages from inappropriate texts. For example, with transcribed spoken language, 'adult' content can be identified during transcription.

Advanced vocabulary

Suitability for children does not only concern adult content, but also the content's level of difficulty. Examples of business and legal language, for example, could potentially include vocabulary and concepts well beyond the expected level of a primary school pupil. Upon completion of the corpus, the CorCenCC project will be in a position to produce the very first frequency lists for Welsh. Plans are underway to implement a search result filter based on frequency, so that corpus-users can match search results to level of ability in Welsh.

'Incorrect' language

Another concern raised by teachers regarding 'suitability' for school children was whether the whole of the corpus content would model 'correct' Welsh. School teaching in Wales is driven by the requirements of the National Curriculum for Wales, and for the Welsh language, even at primary school level, these entail prescriptive expectations in terms of language use.[1] However, it is part of CorCenCC's philosophy to describe the language that exists, rather than to present language data according to prescriptive norms. Corpus data will not be labelled as 'correct' or 'incorrect', and this could potentially limit pedagogical end-users' engagement with the corpus. To ensure that the corpus does fulfil its potential as a learning/teaching resource, the project's solution is to provide guidelines suggesting how Welsh teachers might best filter corpus search results to include only that content which is most likely to reflect the type of language they are trying to model.

Because of CorCenCC's commitment to meeting the needs of end-users, and because those end-users include children, CorCenCC will face challenges not faced by many other corpus projects. However, these challenges also provide opportunities to find user-informed solutions, and to lead the way for future corpora. This paper will look at the pros and cons of some of the possible strategies for tackling 'unsuitable' language (of various kinds) in the corpus, and will illustrate the likely outcomes of the different approaches using Welsh language data. Finally, the paper will propose a 'suitable language policy', which could be adopted not only by CorCenCC but also by future corpora looking to meet the needs of younger corpus-users.

References

DfES (Department for Education and Skills, Welsh Government). 2015. *Welsh second language in the National Curriculum for Wales, Key Stages 2-4*. Cardiff: Welsh Government. Available online: http://learning.gov.wales/resources/browse-all/welsh-second-language-nc/ [Accessed 21-12-2016].

DfES. 2016. *Curriculum for Wales: Programme of Study for Welsh, Key Stages 2-4*. Cardiff: Welsh Government. Available online: http://learning.gov.wales/resources/browse-all/welshnc/ [Accessed 21-12-2016].

[1] For example, for those in Welsh-medium education, Year 6 pupils are expected to be able to 'use… syntax structures and vocabulary… correctly', 'negate sentences correctly', 'mutate correctly after prepositions and pronouns', 'spell correctly' and 'begin to craft their language and write accurately' (DfES 2016, pp. 3-13), and the curriculum for primary school pupils learning Welsh as a second language mentions 'accurate use of a variety of vocabulary, phrases, questions [and] sentence patterns' (DfES 2015, p. 11).