The CorCenCC Crowdsourcing App: A Bespoke Tool for the User-Driven Creation of the National Corpus of Contemporary Welsh

Steven Neale (Cardiff University, UK), Irena Spasić (Cardiff University, UK), Jennifer Needs (Swansea University, UK), Gareth Watkins (Cardiff University, UK), Steve Morris (Swansea University, UK), Tess Fitzpatrick (Swansea University, UK), Lindsay Marshall (Newcastle University, UK) and Dawn Knight (Cardiff University, UK)

Introduction

The CorCenCC project[1] (*Corpws Cenedlaethol Cymraeg Cyfoes* or *National Corpus of Contemprary Welsh* in English; www.corcencc.org) aims to assemble a 10 million-word corpus of the Welsh language across a range of contemporary contexts from spoken, written and e-language sources. In keeping with its contemporary aspect, a key innovation of the project is to facilitate crowdsourced contributions to the corpus, giving Welsh speakers the opportunity to directly involve themselves in the creation of the corpus. This is of vital importance in the Welsh context, in which community pride is strong and for which an open linguistic resource that properly represents the constantly-evolving landscape of contemporary Welsh speakers and the way their language is used is expected to have a wide-reaching impact on the way publishers, policy-makers, the education sector, academic researchers and many more work with Welsh going forward.

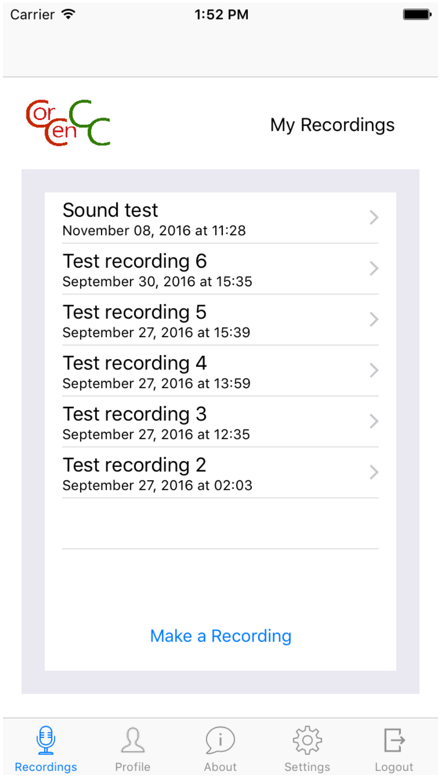
This presentation introduces the CorCenCC Crowdsourcing App, a mobile application designed to facilitate direct contribution of spoken language data to the corpus. Spoken language data will comprise 400,000 of the 10 million word corpus (alongside 400,000 word of written data and 200,000 words of electronic language such as blogs and emails), and app users can contribute directly to this number by recording their Welsh-language narratives (Figures 1 and 2), attaching and editing appropriate metadata to describe the recorded conversations, and uploading them for inclusion in the final corpus. The metadata attached to the recorded conversations includes details about where the recording was made, who else was involved in the recording, and tags that future corpus tools will be able to use to search the data in the final corpus.

Sampling and Ethical Considerations

An important consideration of the crowsdourcing app's design is how the data contributed by users correlates to the sampling frame that has been designed for the CorCenCC project, and specifically to the kinds of Welsh users that need to be reflected and accounted for in the data. New users registering to use the app are required to complete a user profile (Figure 3), which elicits a range of information that has been deemed necessary in order to fully represent the Welsh that is being contributed: speaker’s year of birth, local authority, regions that have been considered to have influenced their Welsh, context(s) that have influenced the development of their Welsh, and how a speaker would describe their ability to use Welsh, among others. This information about the contributors is vital in enabling

[1] 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).

recorded conversations – and eventual Welsh language data – to be attributed to speakers from various contexts and across different social, economic or geographical backgrounds.



Sound test
November 08, 2016 at 11:28

Test recording 6
September 30, 2016 at 15:35

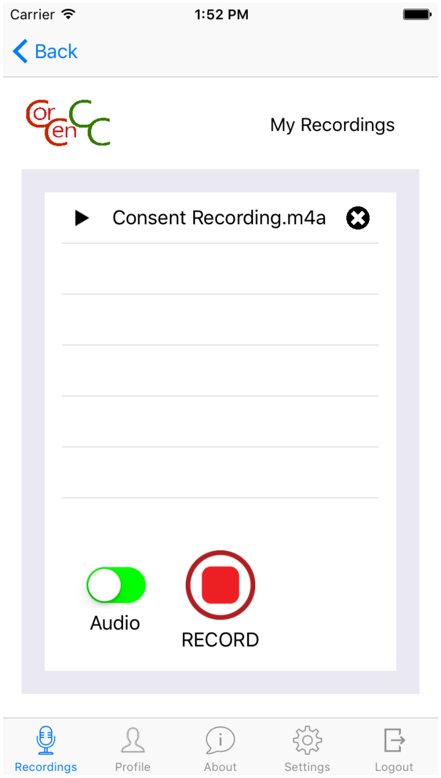
Test recording 5
September 27, 2016 at 15:39

Test recording 4
September 27, 2016 at 13:59

Test recording 3
September 27, 2016 at 12:35

Test recording 2
September 27, 2016 at 02:03

Figure 1. A list of recordings made using the app.



Consent Recording.m4a

Figure 2. The recording screen

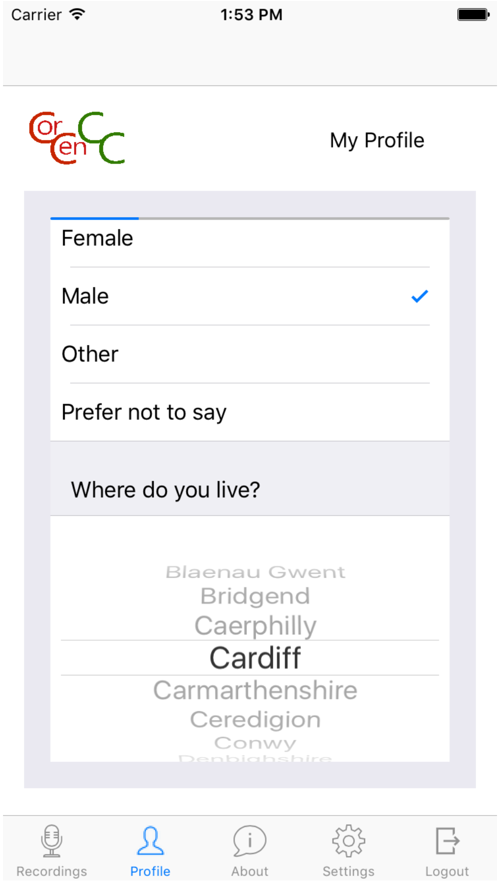
Being able to reflect this information in contributions made via the crowdsourcing application is of huge importance for CorCenCC, whose user-driven remit has been designed to ensure that contributions are truly representative of the Welsh speaking community. Constructing principled corpora involves ensuring that the data they include is balanced, representative and fully documented, and adequate balance and representation can pose particular challenges to maintain – particularly when data arrives in the form of unplanned, spontaneous and ad-hoc contributions, as in the crowdsourcing context. The information collected about contributors by way of carefully constructed user profiles therefore provides the 'fully documented' aspect of the principled corpus design, allowing for any gaps or discrepancies that might surface in the balance and representation of crowdsourced contributions to be identified and properly addressed.

In keeping with the ethical considerations of corpus creation, ensuring that full permissions are sought from and provided by users of the Crowdsourcing App is a vital consideration. After being shown a full description of the project and how data will be used, we ask participants to allow us permission to use their contributions in CorCenCC and for project-related research activities before they are able to register as a user, and also encourage users to record themselves and anybody else involved in a recording verbally stating that they consent for a conversation to be used before they are able to make an actual recording. Additionally, we fully explain that upon upload for inclusion in the corpus, recordings are appropriately anonymised in the same way as a contribution collected by our linguists in the field would be.

Crowdsourcing and Apps in Applied Linguistics

Crowdsourcing methods – where tasks are out*source*d to external contributors (the *crowd*) – are still at a relatively early stage of adoption in applied linguistics, but their successful application in other research areas – including related fields from other disciplines such as Natural Language Processing – are evidence of their merit and their potential impact. Some examples where crowdsourcing methods have been used in applied linguistics have involved the creation and collection of speech and language data (Callison-Burch and Dredze, 2010; Lane et al., 2010), often using tools designed to facilitate outsourcing tasks to external contributors such as the Amazon Mechanical Turk or by completion of bespoke processes such as an online educational game (McGraw et al., 2010). Crowdsourcing methods have also been used in more defined tasks, such as the transcription of language data for speech recognition (Novotney and Callison-Burch, 2010) and the evaluation of translation quality (Callison-Burch, 2009).

Crowdsourcing of language data by way of bespoke applications is in its early stages of adoption, with the *Dialäkt and Voice Äpp* built to collect data to help identify and differentiate between different Swiss German dialects in locations around Switzerland being a noteworthy example (Goldman et al., 2014). However, although larger-scale user-driven corpus construction is as yet unexplored, it is particularly viable and appropriate in the Welsh language context, where community pride is of huge importance and there is a passionate interest in sustaining and 'growing' the language. The CorCenCC Crowdsourcing App can offer a vital resource for harnessing this interest and giving Welsh speakers and learners the opportunity to directly contribute to an open, growing resource that is both *for* them and *by* them.



My Profile

Female

Male ✓

Other

Prefer not to say

Where do you live?

Blaenau Gwent
Bridgend
Caerphilly
Cardiff
Carmarthenshire
Ceredigion
Conwy
Denbighshire

Figure 3. A user profile being filled in

Conclusion and Future Work

The Crowdsourcing App has already gone through a technical testing phase and moving forward we are planning a full human-factors evaluation, using tried-and-tested means such as the System Usability Scale (Brooke, 1996) and a qualitative questionnaire to discern how easy the app is for everyday users. The app is currently available for iOS platforms (iPhone and iPad), with an Android version of the app in development that will widen the potential pool of users (thus, contributors) to CorCenCC in the future. We envisage that the app can play its part in showing that a shift in corpus creation methods is possible, highlighting the important role that crowdsourced contributions can play in constructing user-driven resources and representing diverse groups of language users, in the Welsh context and beyond.

References

Brooke, J. (1996). SUS - A Quick and Dirty Usability Scale. *Usability Evaluation in Industy***,** 189-194.

Callison-Burch, C. and M. Dredze. (2010). Creating speech and language data with amazon's mechanical turk. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, California, 2010. pp. 1-12.

Callison-Burch., C. (2009). Fast, cheap and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP) 2009 (held in conjunction with ACL-IJCNLP)*, Singapore. pp. 286-295.

Goldman, J-P., Leemann, A., Kolly, M-J., Hove, I., Almajai, I., Dellwo, V. and Moran, S. (2014). A crowdsourcing smartphone application for Swiss German: putting language documentation in the hands of the users. Proceedings of LREC 2014, *(Language Resources and Evaluation Conference),* Reykjavik. pp. 3444-3447.

Lane, I., M. Eck, K. Rottmann, and A. Waibel. (2010). Tools for collecting speech corpora via mechanicalturk. In *Proceedings of the North American Chapter of the Association of Computational Linguistics (AACL) Human Language Technologies Workshop 2010*, Los Angeles. pp. 184-187.

McGraw, A. Gruenstein, A. and Sutherland, A. (2009). A self-labeling speech corpus: Collecting spoken words with an online educational game. In *Proceedings of Interspeech 2009*, Brighton. pp. 3031-3034.

Novotney, S. and Callison-Burch, C. (2010). Cheap, fast and good enough: automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10*, Los Angeles, California. pp. 207–215.