The UC QuakeBox Project: Creation of a community-focused research archive

Liam Walsh[1], Jen Hay[12], Derek Bent[1], Liz Grant[1],
Jeanette King[13], Paul Millar[4], Viktoria Papp[12] & Kevin Watson[12]

1 New Zealand Institute of Language, Brain and Behaviour
2 Department of Linguistics, University of Canterbury
3 Aotahi: School of Māori and Indigenous Studies, University of Canterbury
4 Department of English, University of Canterbury

1. Introduction

The University of Canterbury is known internationally for the Origins of New Zealand English (ONZE) corpus (see Gordon et al 2004). ONZE is a large collection of recordings from people born between 1851 and 1984, and it has been widely utilised for linguistic and sociolinguistic research on New Zealand English. The ONZE data is varied. The recordings from the Mobile Unit (MU) are interviews and were collected by members of the NZ Broadcasting service shortly after the Second World War, with the aim of recording stories from New Zealanders outside the main city centres. These were supplemented by interview recordings carried out mainly in the 1990s and now contained in the Intermediate Archive (IA). The final ONZE collection, the Canterbury Corpus, is a set of interviews and word-list recordings carried out by students at the University of Canterbury. Across the ONZE corpora, there are different interviewers, different interview styles and a myriad of different topics discussed. In this paper, we introduce a new corpus – the QuakeBox – where these contexts are much more consistent and comparable across speakers.[1] The

[1] The QuakeBox project is a collaboration between UC CEISMIC and NZILBB at the University of Canterbury, both of whom provided considerable support for the project. Many thanks to all of the NZILBB and UC CEISMIC members who contributed in various ways. It also benefitted greatly from the support of Canterbury Community Trust, PBT Transport and Leighs Construction. We would like to thank Tourism New Zealand for the donation of the QuakeBox itself, and the University of Canterbury Summer Scholarship scheme for funding several student interviewers. The following people worked on the QuakeBox project as interviewers and/or transcribers: Geoffrey Clements, Mark Darbyshire, Andrew Dean, Troy Gillon, Alia Hope-Wilson, Ivan Ignatov, Elizabeth Kivi, Andrew MacFarlane, Henrietta McNeill, Morgana Mountfort-Davies, Michael Peek, Belinda Pflaum, Sally Roome, Maree Shirota, Chelsea Smith, Kirsty Thompson, Kris Vavasour, Lucy-Jane Walsh, Liam Walsh, Yena Wei, Anne Williams. We would also like to thank Scott Lloyd, Robert Fromont and Emma Parnell for their technical and administrative

QuakeBox is a corpus which consists largely of audio and video recordings of monologues about the 2010-2011 Canterbury earthquakes. As such, it represents Canterbury speakers' very recent 'danger of death' experiences (see Labov 2013).

In this paper, we outline the creation and structure of the corpus, including the practical issues involved in storing the data and gaining speakers' informed consent for their audio and video data to be included.

2. Overview of the QuakeBox corpus

In early 2012 the University of Canterbury launched the QuakeBox as part of a collaborative project between the New Zealand Institute of Language, Brain and Behaviour (henceforth NZILBB) and the UC CEISMIC group. The eponymous "QuakeBox" is itself a shipping container which has been converted for use as a transportable recording studio. The objective of the project was to host the QuakeBox at various locations in and around the city of Christchurch, in order to record members of the public telling stories of their experiences of the 2010-2011 Canterbury earthquakes. The transportable recording studio was donated by Tourism New Zealand who had previously outfitted the container and deployed it at sites around New Zealand in 2009 to record tourist impressions of New Zealand as part of their 'Have Your Say' promotion (Tourism New Zealand, 2009). Technical staff at the University of Canterbury refitted and adapted the recording studio for its new purpose. Numerous practical and administrative hurdles were faced in launching the project. These included issues relating to sound proofing, transportation, power and substantial council consenting requirements. These were successfully navigated, and the QuakeBox was finally launched in April 2012.

support. The project also benefitted from the advice or services of Jayne Austin, Gary Busch, Rob Morris, Jessica Petersen, and Rob Stowell. The project would not have been possible without the generous support of the host sites, at: Eastgate Mall, New Brighton Library, Brooklands, Lyttelton, Sumner, the Canterbury A&P show, Westfield Riccarton, and Cashel Mall (Re:Start). This paper has benefitted from the feedback of Alia Hope-Wilson and Lynn Clark.

Figure 1: UC QuakeBox at Eastgate (top left), New Brighton (top right), Re:Start (bottom left) and Lyttelton (bottom right). (all photos Creative Commons NC-ND. Taken by: top photos - Derek Bent; bottom left - Liz Grant; bottom right - Scott Lloyd)

By the end of 2012 the QuakeBox project had recorded 722 stories. Ninety-six of these were recorded in the NZILBB's Observation Lab between 7th December 2011 and 2nd March 2012, with the balance coming from the QuakeBox itself between April and December 2012. During its months in the field, the QuakeBox was stationed at eight locations in the greater Christchurch area. Initially these locations were selected with a focus on areas which had suffered extensive damage as a result of the earthquakes. The QuakeBox's locations, their area within Christchurch, dates on site, and number of stories recorded, are shown in Table 1:

Table 1: QuakeBox recording locations

Site	Relative location in Christchurch	Dates	No. stories recorded
NZILBB Observation Lab	University of Canterbury	7 Dec 2011 – 2 March 2012	96
Cashel Mall (Re:Start)	Christchurch CBD	21 April – 19 May	82
Eastgate Mall	Linwood, eastern Christchurch	22 May- 21 July	197
New Brighton Library	New Brighton, eastern Christchurch	24 July – 6 September	111
Brooklands	north-eastern suburb, greater Christchurch area	9 September – 21 September	24
Lyttelton town centre	south of Christchurch city in greater Christchurch area	2 October – 20 October	69
Sumner village	southeast Christchurch	24 October – 11 November	44
Canterbury A&P Show	Canterbury Agricultural Park, western Christchurch	14 November – 16 November	35
Westfield Riccarton	Riccarton, western Christchurch	20 November – 10 December	64

Participants were asked to complete a survey form covering background information which was, for the most part, specific to events surrounding the earthquakes. This information was in turn used to generate a sizable archive of metadata to aid researchers. Participants were also offered the choice of telling their story in other languages. They could choose whether or not to associate their name with the story they told, and had the option to have only audio data recorded if they did not wish to be filmed. Participants also had a very detailed range of consent options from which to select, granting them a high level of control over exactly how their story would be made available after recording (see Section 3.2).

Participants then told their stories in a private, enclosed booth inside the QuakeBox, optionally accompanied by a member of staff as an interviewer. Those who had consented to being filmed (i.e. the majority) were recorded using a high-definition digital video camera. Audio feeds came from a headset microphone worn by the participant and a ceiling microphone inside the recording booth. The embedded audio in the video file overlays these two inputs, resulting in two identical channels forming a stereo output. However, a separate high-quality audio file was also recorded using a different computer linked to the recording hardware. This file represents the two input signals as separate channels, giving precedence to the headset microphone. The file thus created is more suitable for speech research, as it is not only a higher-fidelity recording, but also it greatly enhances the amplitude of the participant's speech while diminishing the signal from the ceiling microphone. The result is a cleaner output, stripped of most of the background noise found on the video file's audio track.

The QuakeBox received a positive response to its request for stories to be told in languages other than English: in total twelve other languages were recorded, across twenty-five stories. Table 2 details the languages spoken in the QuakeBox corpus, as well as the number of stories told in each language. Many of these participants told their stories twice: once in English, and once in another language.

Table 2: QuakeBox corpus recordings by language

Language	**Number of Stories**
English	697
Mandarin	7
Te Reo Maori	5
Japanese	3
Russian	2
Arabic	1
Cantonese	1
Dutch	1
French	1
German	1
Hungarian	1
Portuguese	1
Punjabi	1
Total	**722**

3. The QuakeBox Corpus

After recording, the files were returned to the NZILBB at the University of Canterbury, where a team of transcribers then created comprehensive time-aligned transcripts of each story in the form of ELAN (Sloetjes & Wittenburg 2008) annotation files. The completed annotation files that resulted were uploaded to a corpus-specific version of LaBB-CAT, the NZILBB's browser-based searchable database (Fromont & Hay, 2012). LaBB-CAT can display a range of data alongside the transcript, and the transcript itself can be exported in different formants (see Figure 2). The export function was used to create a PDF version of the transcript that stripped out all extraneous notations; including e.g., markers denoting pauses between words, instances of noise, phonetic/lexical data entries designed for interaction with LaBB-CAT. By late 2013 the post-production team at the NZILBB had completed transcription of all utilisable stories in the corpus.[2]

The corpus in its entirety contains an estimated 120 hours of recordings. Naturally, consent levels vary across the corpus, but in general participants showed a decidedly candid approach to the QuakeBox. A total of 576 of the 722 stories were flagged by participants for release on the publicly-accessible UC CEISMIC Canterbury Earthquake Digital Archive website. Of those stories remaining, many will be held at the NZILBB solely for purposes related to research.

Stories released for public access can be reached through the UC CEISMIC website,[3] with the QuakeBox collection forming part of the UC QuakeStudies constituent of the UC CEISMIC archives. QuakeBox stories with streamable video and audio are freely and publicly available for viewing there, while certified researchers can request access to restricted content, such as downloadable copies of high-quality video and audio files, ELAN time-aligned transcripts, and HTK aligned phoneme-level alignments (which have not yet been handchecked).[4]

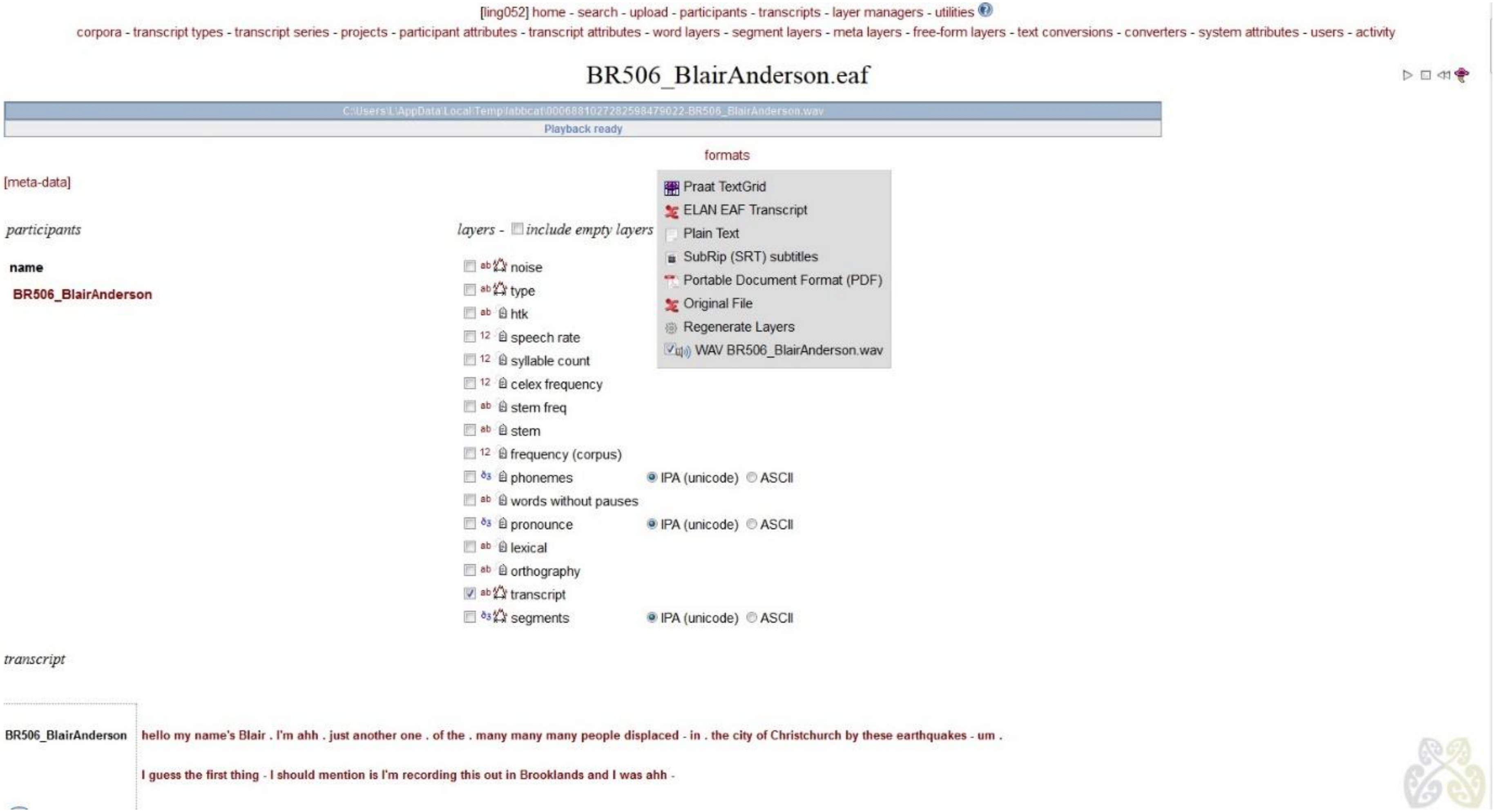
The corpus features a comprehensive demographic cross-section. Participants were asked to provide their ethnicity and age as part of the survey process. The corpus contains a wide range of demographics drawn from the population of Christchurch as a whole, as well as containing a significant amount of input from tourists or visitors – both those who were in Christchurch when the earthquakes happened and those who have come to visit the city since. All age groups are represented in the corpus (from "18-25" through to "85+"), across speakers who collectively are (or were) residents of almost every Christchurch suburb. However, only around 44% of speakers identify themselves as having grown up in Christchurch city or surrounding districts in the North Canterbury area (although the proportion of Christchurch residents was much higher), and almost 25% of all participants grew up outside New Zealand. Thus, the QuakeBox corpus is not just a corpus of recordings of New Zealand English speakers, but of a wide range of speakers who experienced the Canterbury earthquakes.

[2] Certain stories have had to be excluded due to issues relating to either participant consent or file quality; e.g. those with video files that had suffered data loss during the recording process.

[3] See: http://www.ceismic.org.nz/

[4] The collection can be found at: https://quakestudies.canterbury.ac.nz/store/collection/235

Figure 2: An interactive HTML transcript of a QuakeBox story on LaBB-CAT.



[ling052] home - search - upload - participants - transcripts - layer managers - utilities

corpora - transcript types - transcript series - projects - participant attributes - transcript attributes - word layers - segment layers - meta layers - free-form layers - text conversions - converters - system attributes - users - activity

BR506_BlairAnderson.eaf

C:\Users\L\AppData\Local\Temp\labbcat\0006881027282598479022-BR506_BlairAnderson.wav

Playback ready

formats

- Praat TextGrid
- ELAN EAF Transcript
- Plain Text
- SubRip (SRT) subtitles
- Portable Document Format (PDF)
- Original File
- Regenerate Layers
- ☑ WAV BR506_BlairAnderson.wav

[meta-data]

participants

name

BR506_BlairAnderson

layers - ☐ *include empty layers*

- ☐ ab noise
- ☐ ab type
- ☐ ab htk
- ☐ 12 speech rate
- ☐ 12 syllable count
- ☐ 12 celex frequency
- ☐ ab stem freq
- ☐ ab stem
- ☐ 12 frequency (corpus)
- ☐ ðʒ phonemes ● IPA (unicode) ○ ASCII
- ☐ ab words without pauses
- ☐ ðʒ pronounce ● IPA (unicode) ○ ASCII
- ☐ ab lexical
- ☐ ab orthography
- ☑ ab transcript
- ☐ ðʒ segments ● IPA (unicode) ○ ASCII

transcript

BR506_BlairAnderson hello my name's Blair . I'm ahh . just another one . of the . many many many people displaced - in . the city of Christchurch by these earthquakes - um .

I guess the first thing - I should mention is I'm recording this out in Brooklands and I was ahh -

3.1. Participant data

Researchers have a broad collection of metadata to draw on in order to aid their work. Most of the data gathered about participants is personal information relating directly to the earthquakes. As this is a corpus which has resulted from a specific series of events directly affecting, and directly experienced by, an entire local population, it seemed appropriate to define the participant data gathered in accordance with such circumstances. This was important because the archive was not just collected with linguistic research in mind, but also as a general research archive relating to experiences during the Christchurch earthquakes. The meta-data is thus useful for a range of earthquake-related research questions. We were careful not to overly burden our story teller, however, and so some data one might expect to find in a sociolinguistic corpus is missing (probably most notably, questions relating to socio-economic background).

Participants were asked to provide the following personal data:

- Age group (mostly expressed in brackets of 10 years);
- Gender;
- Ethnic group(s);
- Height;
- Where participant grew up;
- Which languages the participant can comfortably speak;
- Where the participant was at the time of the September (2010) earthquake;
- Where the participant was at the time of the February (2011) earthquake;
- Where the participant was at the time of the June (2011) earthquake;
- Where the participant was living prior to the September earthquake;
- Where the participant is living now (i.e., at the time of recording);
- How the house in which they are currently living is zoned;
- Whether they had to move, either temporarily or permanently, because of the earthquakes.

3.2. Consent of participants

Due to the potential for the completed recordings to be used in many different ways, a detailed, multi-tiered consent system was developed. It allowed participants to be extremely specific in tailoring the conditions under which their own individual story would be held in the archives. In the first instance, participants were asked to select from nine different options: four relating to research use and five to public use of their story. Each of these options was further broken down to allow the participant to choose what elements of their story (audio, video, transcript; also images; see below) they wished to make available for that specific purpose.

The consent options relating to research concern the story's transcript, audio, and/or video:

- being made confidentially available to bona-fide researchers based at the University of Canterbury;
- being made confidentially available to bona-fide researchers based at other Universities and institutions;
- being used (in excerpted form) in teaching, public lectures and presentations;
- being played (in excerpted form) to research participants in future research studies.

The public-use options address the participant's willingness to consent that some or all of the media components of their story:

- be made publicly available on the 'UC CEISMIC Canterbury Earthquake Digital Archive repositories' website;
- be made available for use (as transcript, audio, video, and/or images) on any public website via the UC CEISMIC Canterbury Earthquake Digital Archive project;
- be displayed publicly in a museum;
- be broadcast on television or radio;
- be incorporated into other works, such as books, films and artworks.

4. File development and storage

The first stage of post-recording work focused on those stories for which the relevant participants had given their full consent to all options listed on the form. These stories were, in the first instance, re-encoded for use in ELAN, before being transcribed and uploaded into LaBB-CAT.[5] Copies were then delivered to UC CEISMIC, who made the stories publicly available on their website in the form of streamable video with embedded audio. The original recordings remain in a separate archive at the NZILBB.

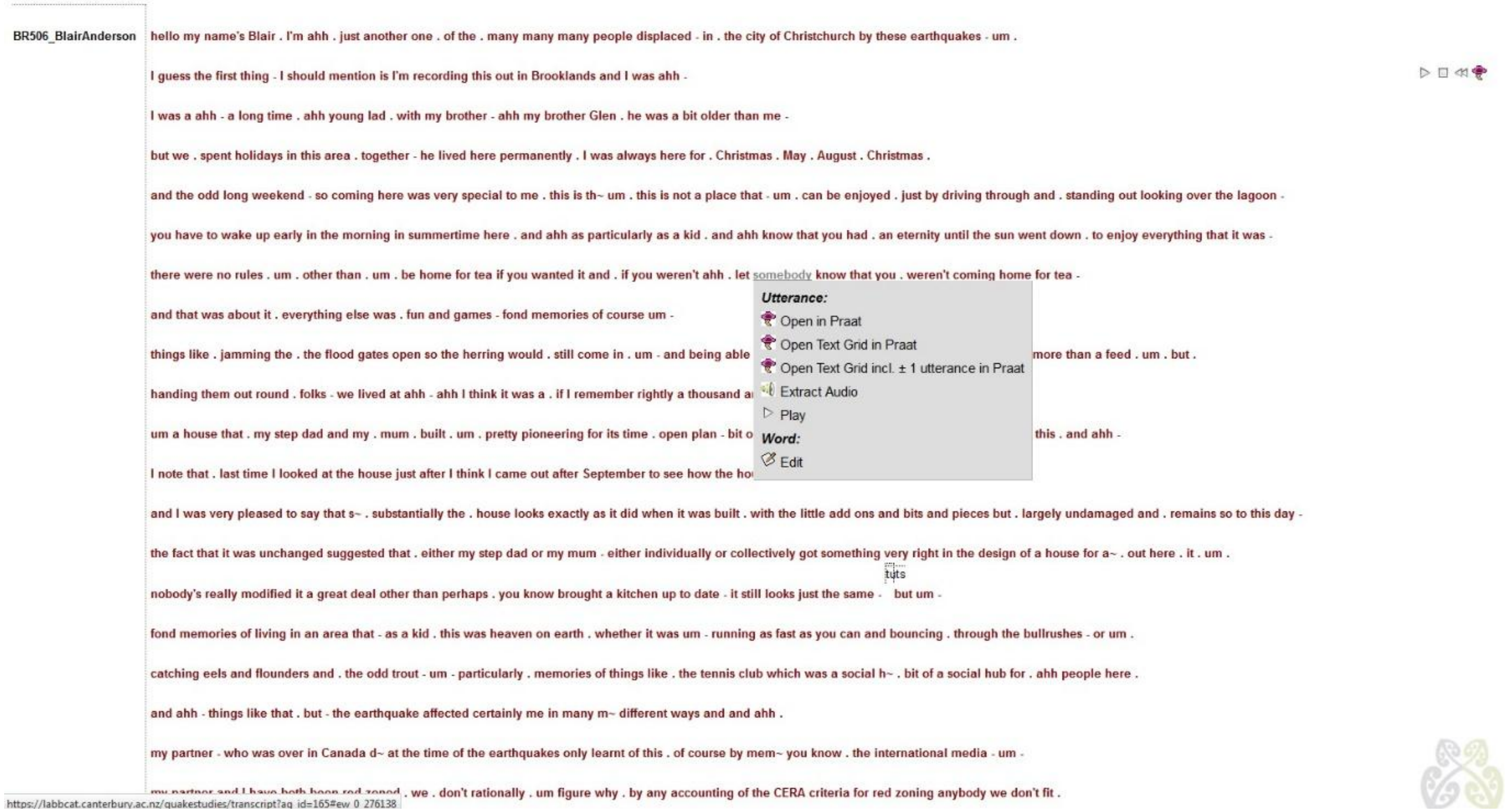
Once work on these initial stories had been completed, the transcription team moved onto preparation of stories for which the participants had not necessarily given consent for the other available options, but had at least consented to all media associated with their story being made publicly available on the UC CEISMIC website. These stories are distributed with different license conditions to reflect the restrictions on their use outside of UC CEISMIC and the NZILBB.

[5] For more detailed information on the structure and functions of LaBB-CAT see Fromont & Hay (2008, 2012).

With the stories uploaded and stored online, the framework of LaBB-CAT makes it easy to interact with the data. Each word within a transcript can be clicked on to open a menu from which one can play audio of the containing utterance, launch the utterance in Praat, view and edit (with or without adjacent utterances) as a Praat textgrid, or export an audio file of the individual utterance (see Figure 3).

It is also possible to use a wide range of search criteria across the archive. The browser-based software enables researchers to perform complex, in-depth searches on the whole QuakeBox corpus. This is of benefit not only to linguists and those researching language-based questions, but also has great potential for academics in other disciplines – such as anthropology, psychology, history, social work etc. – who may wish to investigate e.g. the societal, mental, political, or even commercial impact of the earthquakes, among other things. LaBB-CAT's search functions can be used to perform multi-layered searches using regular expressions, so are of use to anyone looking into topics relating to these natural disasters. At the most basic level, simple keyword searches performed using the orthography or transcript layers will return possible points of interest. The corpus is readily available for use by researchers based at research institutions outside the University of Canterbury, with online access to LaBB-CAT arranged through the NZILBB.

Figure 3: Interacting with a transcript in LaBB-CAT.



BR506_BlairAnderson

hello my name's Blair . I'm ahh . just another one . of the . many many many people displaced - in . the city of Christchurch by these earthquakes - um .

I guess the first thing - I should mention is I'm recording this out in Brooklands and I was ahh -

I was a ahh - a long time . ahh young lad . with my brother - ahh my brother Glen . he was a bit older than me -

but we . spent holidays in this area . together - he lived here permanently . I was always here for . Christmas . May . August . Christmas .

and the odd long weekend - so coming here was very special to me . this is th~ um . this is not a place that - um . can be enjoyed . just by driving through and . standing out looking over the lagoon -

you have to wake up early in the morning in summertime here . and ahh as particularly as a kid . and ahh know that you had . an eternity until the sun went down . to enjoy everything that it was -

there were no rules . um . other than . um . be home for tea if you wanted it and . if you weren't ahh . let somebody know that you . weren't coming home for tea -

and that was about it . everything else was . fun and games - fond memories of course um -

things like . jamming the . the flood gates open so the herring would . still come in . um - and being able more than a feed . um . but .

handing them out round . folks - we lived at ahh - ahh I think it was a . if I remember rightly a thousand a

um a house that . my step dad and my . mum . built . um . pretty pioneering for its time . open plan - bit o this . and ahh -

I note that . last time I looked at the house just after I think I came out after September to see how the ho

and I was very pleased to say that s~ . substantially the . house looks exactly as it did when it was built . with the little add ons and bits and pieces but . largely undamaged and . remains so to this day -

the fact that it was unchanged suggested that . either my step dad or my mum - either individually or collectively got something very right in the design of a house for a~ . out here . it . um .

tuts

nobody's really modified it a great deal other than perhaps . you know brought a kitchen up to date - it still looks just the same - but um -

fond memories of living in an area that - as a kid . this was heaven on earth . whether it was um - running as fast as you can and bouncing . through the bullrushes - or um .

catching eels and flounders and . the odd trout - um - particularly . memories of things like . the tennis club which was a social h~ . bit of a social hub for . ahh people here .

and ahh - things like that . but - the earthquake affected certainly me in many m~ different ways and and ahh .

my partner - who was over in Canada d~ at the time of the earthquakes only learnt of this . of course by mem~ you know . the international media - um -

my partner and I have both been red zoned . we . don't rationally . um figure why . by any accounting of the CERA criteria for red zoning anybody we don't fit .

https://labbcat.canterbury.ac.nz/quakestudies/transcript?ag_id=165#ew_0_276138

5. Characteristics of the QuakeBox corpus

The QuakeBox corpus, unlike other corpora housed at the NZILBB, is a collection that relates to a very specific set of subject matter. Seldom is a corpus so focused on the many possible experiences that may derive from a single, multi-faceted event, like an earthquake, that affects an entire city. This is one of the factors that makes the QuakeBox archive so valuable as a research corpus: the focus on the same narrow subset of topics by every participant allows, to a certain extent, a researcher to control for a number of significant variables faced when working with less uniform corpora. Such variations may be caused by differences in content, enthusiasm, spontaneous storytelling ability, relevance of events, reasons for the recordings taking place, etc. This last point of difference is easily accounted for in the QuakeBox recordings: participants were all members of the public who approached the QuakeBox and gladly volunteered their stories, contrasting with participants in a lab study who may be instructed to "speak spontaneously" with little preparation or motivation to do so. In the wake of the earthquakes however, people's experiences of the events were obviously a prolific topic of conversation among the population of Christchurch. Thus, certain QuakeBox participants are likely to have been telling stories which, in a manner of speaking, were to some extent "rehearsed".

Significantly, the QuakeBox stories are almost invariably monologues – with a small number of exceptions, these stories were recorded without significant interaction between participant and an interviewer or other person. While interviewers are frequently present during the storytelling (at the choosing of the participant), they seldom contribute much to the story, preferring instead to let the story to be told without extensive dialogue. Since most of the stories also have accompanying video (in which, again, an interviewer may or may not be present), the archive is also of benefit to those researching gestures in speech.

In addition, the QuakeBox corpus serves as another important archive in the NZILBB's ongoing study of the development of New Zealand English – a localised population discussing a shared experience will always have great value as a historical record of contemporaneous language use. In this regard, the data so gathered will serve to act as an important complement to the already-comprehensive ONZE corpora housed at the NZILBB (see Fromont & Hay, 2008).

6. Conclusion

The QuakeBox project has been a successful venture, with many Christchurch residents contributing to the archive. For some it was an opportunity to speak openly about the earthquakes for the first time. The creation of a digital archive focused on the Canterbury earthquakes from the perspective of individual people has significant historical value. As a permanent record of the mindset of Christchurch people in the time after these disastrous events, the QuakeBox project may serve to inform people in other parts of New Zealand, as well as in other countries. The publicly-accessible nature of the project recordings may well lead to improved understanding of earthquakes and their aftermath. Moreover, the potential for extensive and beneficial research is huge, since academics of all disciplines will have the chance to examine local accounts of a truly extraordinary event.

The Canterbury earthquakes have inflicted radical changes on the city of Christchurch and surrounding towns. The UC QuakeBox project ensures that the stories of those affected will endure, enhancing knowledge of earthquakes and their consequences for universal future benefit. Linguistically, the archive is a valuable resource, constituting what is probably the largest archive of high quality audio and video 'danger of death' stories (Labov 2013) available anywhere.

References

Fromont, Robert and Jennifer Hay. 2008. ONZE Miner: the development of a browser-based research tool. Corpora 3(2), 173–193.

Fromont, Robert and Jennifer Hay. 2012. LaBB-CAT: an Annotation Store. In Proceedings of Australasian Language Technology Association Workshop, 113–117.

Gordon, E., Campbell, L., Hay, J., Maclagan, M., Sudbury, A. and Trudgill, P. (2004) New Zealand English: Its Origins and Evolution. Cambridge: Cambridge University Press.

Labov, William. 2013. The Language of Life and Death: The Transformation of Experience in Oral Narrative. Cambridge: CUP.

Sloetjes, H & P. Wittenburg. 2008. Annotation by category - ELAN and ISO DCR, in proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Tourism New Zealand. 2009. http://www.tourismnewzealand.com/tourism-news-and-insights/latest-tourism-news/2009/10/end-of-the-road-for-mobile-recording-studio/#