

## Exercise 1 Exploring LaBB-CAT

LaBB-CAT is a speech/language corpus management system that:

- stores transcripts with audio/video
  - supporting a variety of formats
  - and the definition of speech elicitation tasks;
- allows the addition of different layers of annotation, which can
  - be manual or automatic, and
  - have different granularities, from topic tagging to individual phones;
- supports forced alignment to phone level using a speech recognition toolkit called “HTK”;
- allows cross-layer regular-expression search;
- search results are exportable to CSV for further analysis;
- batch acoustic measurement of segments using Praat is also supported, and
- transcripts and fragments of them are exportable in a variety of formats.

In this worksheet you will start exploring a demo LaBB-CAT corpus, to get a general idea of how to find your way around LaBB-CAT and how the language data is presented.

The demo corpus contains a collection of videos of people telling stories about their experiences during the earthquakes that struck Canterbury during 2010 and 2011. They have been orthographically transcribed using a tool called ELAN, so they have been time aligned to the utterance level; i.e. the start and end time of each line in the transcript has been manually synchronized with the recording. The ELAN transcripts, and their video and audio files, have been uploaded into LaBB-CAT.


LaBB-CAT is a browser-based system so the first thing to do is access it with your web browser. Generally, any modern browser should be fine (although some features you’ll see in later worksheets are only supported by Mozilla Firefox or Google Chrome).

1. In your web browser, type in the following URL:  
*<https://labbcats.canterbury.ac.nz/demo>*  
NB: ensure you enter ‘https’ not ‘http’  
You will be asked for a username and password.
2. The username is **demo** and the password is **demo**  
The very first time you access LaBB-CAT, you will see its licence agreement.
3. Scroll to the bottom of the page and click *I Agree* to continue.  
You will see a page called “LaBB-CAT Demo” which has a menu of links along the top and a number of icons. Below the icons is some information about the corpus. This is the LaBB-CAT home page.
4. Click the *where do I start?* icon on the left.  
The help page that pops up includes a brief description of LaBB-CAT and some tips for navigation and getting more information.
5. Read through the page, and then close the browser tab to return to the home page.

## 1 Transcripts

You can browse the transcripts and recordings in the data store directly.

6. On the LaBB-CAT home page, click the *transcripts* link on the menu at the top of the page.  
You will see a list of transcripts. Only the first 20 transcripts shown; there are links to other pages of the list at the bottom of the page.
7. Click the name of the first transcript.  
You will see a page with transcript text, and the video appears in the top right corner of the page.

8. Click the play button.  
As the video plays, you will see the current utterance highlighted in the transcript. You will also see that the current utterance appears as closed captions in the video. You can use the video controls as normal, including the *full-screen* button in the bottom right, to make the video occupy the whole screen.
9. Try clicking the magnifying-glass icon  below the video, to see what it does.
10. Pause the recording.
11. Click one of the utterances further down the transcript.  
You will see a menu appear with various options, including a ‘Play’ option at the bottom of the menu.
12. Click the ‘Play’ option on the menu.  
You will see that playback starts at that utterance. Playback will stop when the participant finishes the utterance.
13. Click on the *formats* link under the title.  
You will see a menu, which includes various formats for exporting the transcript.
14. Select ‘Text Only’
15. Click *Convert*
16. Save the resulting file on your desktop, and then open it.  
You will see the transcript in plain-text form.
17. Back in LaBB-CAT, click the browser’s *back* button to return to the transcript.
18. If you have Praat installed on your computer, click the *formats* link, and select the ‘Praat Text Grid’ option. Save the resulting file on your desktop, and then open it with Praat.  
You will see that the TextGrid has various tiers, one for whole utterances, and one for individual words.
19. Back on the transcript page in LaBB-CAT, you will see a link labelled *[meta-data]*; click it.  
You will see some meta-data about the transcript and recording.  
LaBB-CAT attaches meta-data both to transcripts (called ‘transcript attributes’), and also to participants (‘participant attributes’).
20. Below the transcript attributes is the name of the participant. Click their name.  
You will see a page with the participant attributes, and a list of the recordings they appear in. In this case, they appear in only one recording; if you were to click the name of the recording, you would be taken back to the transcript page you’ve just seen.

## 2 Orthographic Search

Searching is a two-step process: first you select which participants you want to search, using their participant attributes. And then you specify the pattern you want to search for.

If we were interested only in monolingual speakers, for example, we would filter out those that speak various language by setting the attribute values appropriately on the filter page.

21. On the menu at the top of the page, there’s a *search* option. Click it.  
You will see a page called “filter”.
22. Select ‘English’ from the *Languages:* dropdown box.  
You will see a list of the participants who only speak English.  
Notice that each participant has a check-box; if we wanted to, we could select specific participants from the list by checking/unchecking the boxes. (But in this case, let’s search all of them, so leave all the boxes ticked.)
23. Select ‘Female’ from the *Gender:* dropdown box.  
You will see the list has been further narrowed down to on female monolingual English speakers.
24. Click the *Layered Search* button at the bottom of the list.  
You will see a page that lists the speakers on the top left, a number of annotation layers on the top right, and has a ‘search matrix’ below. (It doesn’t look much like a matrix yet, as it only includes the ‘orthography’ layer, but we will be adding rows and columns later on)


25. In the “orthography” box enter the word **quake** and click the *Search* button at the bottom.  
A progress bar will appear, and then shortly after that, a new window will open, which has a list of search results in it. Your browser’s popup-blocker might prevent the results page from opening – you can fix that either by allowing the popups in your browser, or by clicking the *Display results* link that appears after the search finishes.  
You will see a list of hits, highlighted within their immediate context in the transcript, grouped by transcript.
26. Click the first result.  
You will see the transcript page, as we saw earlier, but with each match from the search highlighted.
27. Close the transcript tab.  
You’ll be taken back to the search results page.

The search results page has a number of other options as well.

28. Each result line has a ticked checkbox next to it. Untick the “[select all  $n$  results]” checkbox, and then tick a handful of results in the list.  
*Tip:* You can select a group of matches by ticking the first one, and then holding down the < **Shift** > key while ticking the last one.
29. Click the *Extract Audio* button.
30. Save and open the resulting zip file.  
You’ll see that the files are systematically named to include:
  - the result number
  - the name of the transcript
  - the start and end time of the extracted utterance
31. If you also have Praat installed on your computer, go back to the results page and select ‘Praat TextGrid’ from the dropdown list next to the *Convert* button, click *Convert*, and save and open the resulting zip file.  
You’ll see that the TextGrid names match the audio file names in the previous zip file.
32. Back on the results page, click the *CSV Export* button.
33. Save the resulting file, and open it.  
You may have to specify some import options, in which case it may be handy to know that the field separator is comma, and the fields are quoted by speech marks.  
*Tip:* If you’re using Microsoft Excel and you find it doesn’t open all the columns correctly:
  - (a) Create a new workbook in Excel.
  - (b) Click the ‘Data’ tab.
  - (c) On the “Get External Data” ribbon click ‘From Text’.
  - (d) Select the CSV file you downloaded.
  - (e) Select ‘Delimited’ and click *Next*.
  - (f) Ensure ‘Comma’ is the only delimiter ticked and click *Next*.
  - (g) Click *Finish* and then *OK*.

You will see a spreadsheet with one line per selected result, and various columns containing information about the speaker, the corpus, the match line and word, and a URL to the interactive transcript for the match.

With this spreadsheet, you can work ‘offline’ with the results, tagging them, computing statistics in Excel, R, or any other program that can work with CSV files. We’ll look at a few more uses for the CSV results files later...

34. Close the CSV file.  
You’ll be taken back to the results page.  
 If in doubt about a search option, try the online help page.
35. Close the search results tab.  
You’ll be taken back to the search matrix page.

You can also search across multiple words, and search for patterns as well as exact spellings.

For example, let's say you want to investigate how the pronunciation of the word 'the' changes when the following word starts with a vowel. You can search for this pattern using the search form:

36. Click the *search* option on the menu.

37. Click the *Search Everyone* button.

38. Search for the word **the**.

You will see that there are lots of results, including many where 'the' is followed by a word that starts with a consonant.

39. Close the search results tab.

You'll be taken back to the search matrix page.

40. On the search page, underneath the list of layers, there's a box with the number 1 in it. Change the number to 2 and click *Set Search Matrix*.

Now you will see that our search matrix is one layer high by two words wide.

41. In the second, empty "orthography" box, enter: `[aeiou].*`

This is a 'regular expression' that allows you to identify a pattern, with the following parts:

- any vowel ('[aeiou]')
- followed anything at all – '.' in a regular expression means 'any character', and '\*' means 'zero or more of the previous thing', so '.\*' means 'zero or more characters'

42. Click *Search*

You will see that the results include only instances where the word that follows 'the' starts with a vowel. You might see some 'false-positives' like "the one", where the spelling includes a vowel "o" but the pronunciation is actually a consonant /w/, but by and large, they'll be words that are pronounced with a vowel at the beginning.

43. See if you can create a search for all words ending in 'ing'

You can get more information about regular expressions by using the online help back on the search page, and also by clicking the *regular expressions* link above the search matrix.

In this worksheet you have seen that:

- LaBB-CAT is a repository for recordings and their transcripts;
- Meta-data can be attached to transcripts (transcript attributes) and to participants (participant attributes);
- You can search the texts of the transcripts;
- You can filter the search results on the basis of meta-data;
- You can search for patterns as well as exact spelling, by using regular expressions;