1 - Exploration

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", or the Montreal Forced Aligner, or the WebMAUS service provided by BAS Web Services;
- 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://labbcat.canterbury.ac.nz/demo
   You will be asked for a username and password.

> **!** Important
>
> If typing this out manually, ensure you enter 'https' not 'http'

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.

There are two main ways to use LaBB-CAT:

- easy exploration and plain text search (this worksheet)
- layered filtering and search (the following worksheets)

Easy exploration

Using the 'easy' method for exploring LaBB-CAT is simple, as annotation layers are largely ignored; each transcript is essentially treated as a plain text that you can search and display based on ordinary orthographic spelling.

1. On the LaBB-CAT home page, click the *explore* icon to access the easy exploration pages.
   You will see a similar home page, with *Browse* icon, *Easy Search*, and *Layered Search* icons.
2. Click the *Browse* icon.
   You will see a page that lists the collections of recordings (or 'corpora') in the LaBB-CAT database. Each corpus contains a number of recordings.
3. Click the first corpus listed.
   You will see a page that lists the first 20 recordings in the corpus. The recording names are on the left, followed by some meta data (called 'participant attributes') about the participant in the recording. At the bottom is a list of pages, so you can access further recordings in the corpus.
4. Click the name of the first recording.
   You will see a page with transcript text, and the video appears in the top right corner of the page.
5. Hover the mouse over the video.
   The video pane grows larger.
6. Press 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.
7. Pause the recording.
8. Move the mouse over one of the utterances further down the transcript.
   You will notice that the video pane shrinks again, and that the mouse pointer becomes a *play* button.

9. Click the utterance.
   You will see that playback starts at that utterance. Playback will stop when the participant finishes the utterance.
10. At the top of the page, you will see a tab button labelled *General*; 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').
11. 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.

Basic search

1. On the menu at the top of the page, there's a *Search* option. Click it.
   You will see a search form with a "Text" search box at the top, and options for meta data below.
2. In the "Text" box enter the word `quake` and press the *Search* button at the bottom.
   You will see a list of hits, with the name of the transcript on the left and the matched word on the right, highlighted within its immediate context.
3. Click the *Search again* link (or the *Search* link on the menu at the top)
   You'll see that the search form remembers the last search text.
4. Select 'Female' from the *Gender* drop-down box, leave the word `quake` in the "Text" box, and click *Search*.
   This time the results are narrowed down to only female participants.
5. Click the first result.
   You will see the transcript page, as we saw earlier, but with each match from the search highlighted.

Regular Expressions

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:

6. Click the *Search* option on the menu.
7. 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.
8. Go back to the *Search* page.
9. Now search for: `the [aeiou].*`
   This is a 'regular expression' that allows you to identify a pattern, with the following parts:

- the word ‘the’
- followed by a space
- followed by 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’

  You will see that the results include only instances where the word that follows ‘the’ starts with a vowel.

10. See if you can create a search for all words ending in ‘ing’

In this worksheet you have seen that:

- LaBB-CAT is a repository for recordings and their transcripts;
- Transcripts are grouped together into corpora;
- 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.