3. Uploading Data

In this exercise you will:

1. Upload a transcript manually
2. Upload many transcripts at once using the batch uploader
3. Import participant data from a CSV file
4. Define a speech elicitation task for gathering data

After this you will have a small corpus in your LaBB-CAT database.

Before you start, download and unzip QuakeStories.zip so you've got the demonstration data for uploading to your corpus.

Manual Upload

(1) In LaBB-CAT, select the *transcripts* option in the menu.
(2) Press the *Upload Transcript* icon.

You will see a page with some options to select on the top left, buttons on the top right, and in a middle, a rectangle with a dashed border; this is the 'upload queue', which lists files we want to upload.

(3) In the top left corner of the 'upload queue' rectangle, there's a *Choose Files* button; press it, and select the file in the "QuakeStories" folder called `BR178LK_MargaretSpencer.eaf`

You will see that the transcript file is listed in the 'upload queue'. We want to upload not only the transcript, but also its associated media files. Each transcript has an audio file and a video file, and you want to upload both.

(4) Press *Choose Files* button again, and in the same "QuakeStories" folder click the file called `BR178LK_MargaretSpencer.mp4`, then hold down the `Ctrl` (linux), `Ctrl` (windows), `command` (mac) key on your keyboard and click the file called `BR178LK_MargaretSpencer.wav` so that both files are selected.
(5) Then press *Open* (or in some browsers the button to select files is labelled *Upload*). You will see that next to the `BR178LK_MargaretSpencer.eaf` transcriopt, under the *Media* heading, two media types are now show; "mp4" and "wav".
(6) To the right of this, ensure the *Corpus* option is *QB*
(7) Also ensure the *Type* option is *interview*

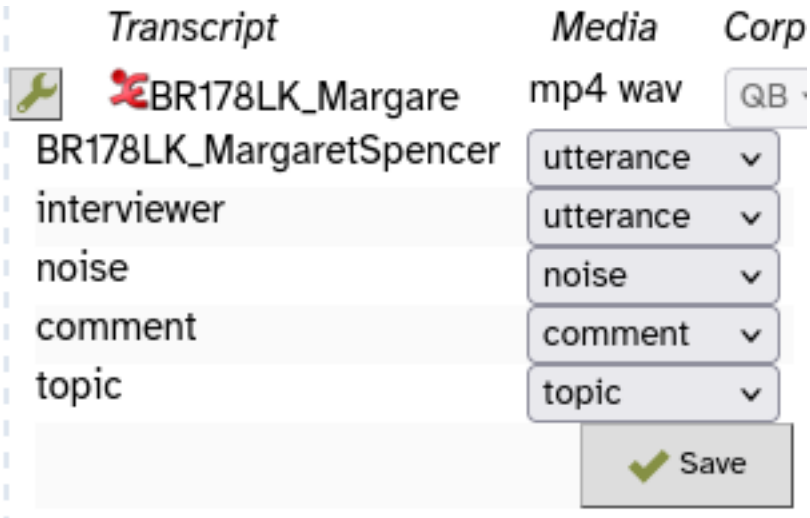
(8) Leave the other options with the default values and press the *Upload* button above.
You will see that, on the right, a progress bar shows 50% progress, and below the transcript in the upload queue, a number of options have appeared.

Each ELAN transcript has a number of Tiers defined in it:

- one for the participant's utterances,
- another for an 'interviewer' if there is one,›
- one for noise annotations,
- one for transcriber comments, and
- one for topic annotations.

Each tier must be mapped to a LaBB-CAT annotation layer.

LaBB-CAT has analysed the structure of the ELAN transcript and pre-selected some default options for layer mappings. For this data, these defaults are correct, so you needn't change anything.



(9) Press *Save* to continue.
You will see that the progress bar on the right continues, and after a short delay, the progress is complete, and the *Status* is listed as "Finished."

(10) The name of the transcript on the left, *BR178LK_MargaretSpencer.eaf*, is now a link. Click it.

You will see a page with transcript text, and the video appears in the top right corner of the page.

(11) Press the play button on the video.
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 to make the video occupy the whole screen.

(12) Pause the recording.

(13) Click one of the transcript lines further down the transcript.
A menu will appear.

(14) Select the 'Play' option on the menu.
You will see that playback starts at that line. Playback will stop when the participant finishes the utterance.

(15) Select the *Formats* tab at the top of the transcript.
You will see a list of formats for exporting the transcript to.

(16) Select *Plain Text Document*

(17) Save the resulting file and then open it.
You will see the transcript in plain-text form.

(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 (or two if there are two speakers), and one for individual words (or two if there are two speakers).

(19) Back on the transcript page, select the *Attributes* tab at the top right.
This will display the attributes for the transcript (some of the attribute values are not set because the information was not in the .eaf transcript file)

(20) Now select the *Participants* tab on the top right.
This will list both participants in the recording, the main participant, and the interviewer.

(21) Click *BR178LK_MargaretSpencer*.
This will display the participant meta-data. There's not much here yet; we will be adding participant attributes soon. However, we can at least set the participant's gender now.

(22) *BR178LK_MargaretSpencer* is an 'English'-speaking 'Female' who is between '66 and 75 years' old, who grew up in 'Christchurch', in the 'North Canterbury' region of 'New Zealand'.
Set her attributes to reflect that, and press *Save*.

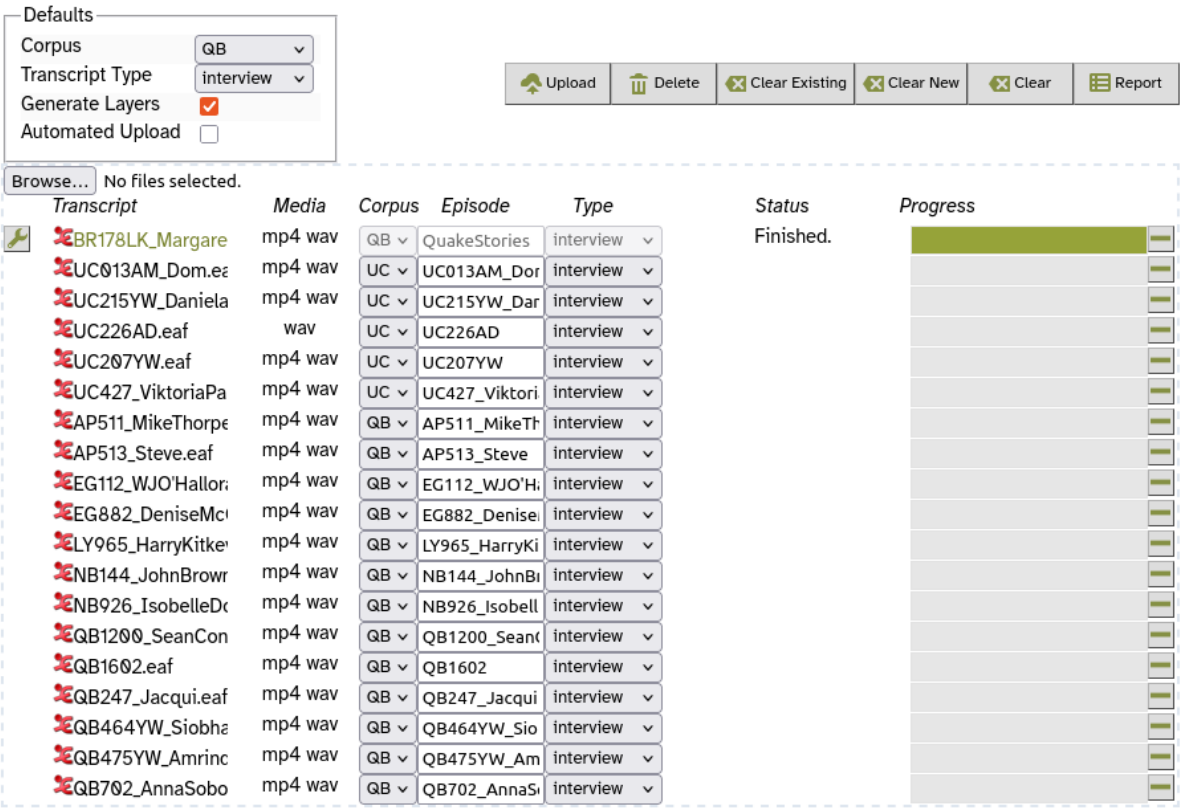
You have now manually uploaded one transcript, checked the ELAN-tier to LaBB-CAT layer mappings and manually specified the meta-data for one participant.

Automated Upload

If you already have a collection of transcripts and media files (which we have for these exercises), and they are systematically organized (which they are), you may be able to save some manual uploading work by uploading them using the 'automated upload' option.

(23) When you clicked the name of the transcript to open it after uploading, a new browser tab was opened. Close that tab now to take you back to the upload queue.
(24) Most of the transcripts we are going to upload are monologues, so in the *Defaults* box on the top left, set *Transcript Type* to *monologue*.
(25) Open Windows Explorer or Finder, and navigate to the LaBB-CAT Workshop data folder.
(26) Drag the folder called "QuakeStories", and drop it on to LaBB-CAT, on to the upload queue area below the buttons (the rectangle with the dashed border).

The upload queue will now contain a longer list of transcripts. Each transcript should have a value filled in for each column - *Transcript*, *Media*, *Corpus*, *Episode*, and *Type*.



(27) The first transcript, *BR178LK_MargaretSpencer.eaf*, has already been uploaded, and we don't want to upload it again. Remove it from the list by using the button on the right hand side of that row.

When we uploaded manually before, we saw a list of ELAN tiers and their correspondences to LaBB-CAT layers. The options had default values, but we had to manually confirm the choices that LaBB-CAT had made about how to interpret the ELAN tiers.

The *Automated Upload* option allows LaBB-CAT to automatically use these default selections, instead of asking us to manually confirm them for every transcript. For this corpus, the default options that LaBB-CAT automatically selects will always be correct.

(28) Tick the *Automated Upload* checkbox in the *Defaults* box on the top left.
(29) Press the *Upload* button above the list.
You will see that in the *Status* column, the text changes to "Uploading..." for the first transcript. The progress bar progresses, and once it's complete, the next transcript changes to "Transferring", and so on.

> **Tip**
>
> While the files are uploading, click the online help link at the top of the page to the right of the menu and check preconditions for uploading, and other functions the upload page can perform.

(30) Once the uploader is finished, you will receive a CSV 'upload report' file that lists the files you uploaded and their upload status. (If there had been any problems with the upload, the resulting error messages would be included in this report for following up.)
(31) You can verify that all the transcripts are there by selecting the *transcripts* option on the menu in LaBB-CAT.
You should see a list of twenty transcripts.
(32) Use the *Transcript* box to find `UC013AM_Dom.eaf`
(you can type just part of the name if you like).
(33) press the *Attributes* icon for *UC013AM_Dom.eaf*
(on the far right of the row).
(34) Change *Transcript type* to *interview* and press *Save*.
(35) Similarly, the following transcripts are interviews, so change their type accordingly

- UC215YW_DanielaMaoate-Cox.eaf
- UC226AD.eaf

Participant Data Import

The transcripts are now in the database, but the meta-data for the participants hasn't been set yet (because it's not contained in the ELAN files). We could manually add this for each speaker, but fortunately we have it stored in a spreadsheet (actually, a CSV text file) that we can upload in one go.

(36) In LaBB-CAT, select the *participants* option on the menu.
(37) Press the *Upload Participant Data* icon at the bottom.

(38) Press *Choose File*, and select the file in the LaBB-CAT Exercises data folder called *participants.csv*
(39) Press *Upload*
(40) You will now see a list of the columns from the spreadsheet.
Firstly, ensure that the *Participant identity column* is set to *name*. This ensures that the "name" column in the spreadsheet will be used to match names of participants in the LaBB-CAT database.
(41) Below that is listed each column from the spreadsheet, with an arrow pointing to a dropdown box. The box contains various options, including each of the participant attributes set up in LaBB-CAT, an *ignore* option, and *create a new attribute* option.
Most likely, the correct options are already selected, as we've already set up the correct participant attributes, but just check that they are as follows:

- The CSV column **name:** → *ignore* because it's the *Participant Identity Column* identified above
- The CSV column **gender:** → the *Gender* LaBB-CAT attribute
- The CSV column **ageCategory:** → the *Age* LaBB-CAT attribute
- The CSV column **ethnicity:** → the *Ethnicity* LaBB-CAT attribute
- The CSV column **grewUup:** → the *grewUp* LaBB-CAT attribute
- The CSV column **grewUpRegion:** → the *grewUpRegion* LaBB-CAT attribute
- The CSV column **grewUpTown:** → the *grewUpTown* LaBB-CAT attribute
- The CSV column **languagesSpoken:** → the *languagesSpoken* LaBB-CAT attribute

(42) Press *import*.
You should see a page with information about the import, including the columns that were ignored, and the number of participants that were added.

To check the participant attributes really are now set:

(43) Select the *participants* option on the menu. You will see a list of speakers, and page links at the bottom.

The page also includes participant attribute values where they are known.

You can also filter the list by these values, using the column headings above the list:

(44) Under *Gender*, select the *F* option.
The page now lists only those with 'Female' set for the *Gender* attribute.

Elicitation Tasks

LaBB-CAT can also make recordings of speech directly from the browser.

Let's suppose you want to record a number of participants reading lists of words. You can define an 'Elicitation Task' that includes a series of steps, one for each set of words you want participants to read.

First we're going to create a corpus to receive our recordings, and a transcript type to mark the recordings as word lists …

(45) In LaBB-CAT, select the *corpora* option on the menu.
(46) Add a corpus called *CC* with a description *Canterbury Corpus*.
(47) Click the *transcript types* option on the menu.
(48) Add a transcript type called *wordlist*.

Now we'll create the elicitation task, which defines what prompts and texts the participant sees during the task.

(49) Click the *elicitation tasks* option on the menu. The page you see is a list of elicitation tasks defined, which is currently empty.
(50) Fill in the blank form with the following details:

- **ID:** `nze-wordlist`
- **description:** `New Zealand English Word List`
- **corpus:** *CC* (the corpus you just created)
- **transcript type:** *wordlist* (the transcript type you just created)
- **preamble:** "In this task your speech will be recorded. Please ensure you're in a quiet place."
  This is the first text the participant sees when they access the task, before giving consent or going through the steps.
- **consent:** "I give consent for the use of my speech data for this research."
  This is the text of the participant's consent for their participation and the use of their data. Before starting the task steps, they must 'sign' this consent by typing their name in a box at the bottom. The text, with their name and the date incorporated, with be made into a PDF file which is uploaded with their recordings, and is made available for them to download.

> **Note**
>
> For both the preamble and the consent form, you can format the text with bold, italic, and underlined text, etc. by using the controls above the text area.

> **Tip**
>
> Check the online help on this page for further details about settings and important information about browser limitations.

(51) Press *New* to add the task.
(52) Press *Define Steps*.

On this page you are going to add steps for the task. The first step, called “Welcome”, has already been added, and we’ll use it for giving the participant some detailed instructions about what follows. We'll add a series of steps after the “Welcome” step, one for each group of words we want the participant to read.

(53) The form you can see defines the details of the first “Welcome” step.
Check the online help on this page for further details about this page and the options on it.

(54) Close the online help page to return to the “define elicitation steps” page.

(55) Fill in the following details:

- **Show:** *Always*
- **Countdown Seconds:** 0
- **Title:** Instructions
- **Prompt:** Please read aloud the following sets of words. Press “Next” after each set.
- **Elicit:** *Nothing*
- **Transcript:** (leave this box blank)
- **Image/Video:** *no image/video*

Next we’ll define what demographic information we will ask each participant before they start recording. In this case, we will ask for their gender and what languages they speak.

(56) Press the button to add a new step.
(57) Fill in the following details:

- **Show:** *Always*
- **Countdown Seconds:** 0
- **Title:** Languages
- **Prompt:** What languages do you speak?
- **Elicit:** *Attribute Value*
- **Attribute:** *participant_languagesSpoken*
- **Image/Video:** *no image/video*

(58) Press the button to add a new step
(59) Fill in the following details:

- **Show:** *Always*
- **Countdown Seconds:** 0
- **Title:** Gender
- **Prompt:** What is your gender?
- **Elicit:** *Attribute Value*
- **Attribute:** *participant_gender*
- **Image/Video:** *no image/video*

Now we can defined some prompts for them to read aloud.

(60) Press the button to add a new step
(61) Fill in the following details:

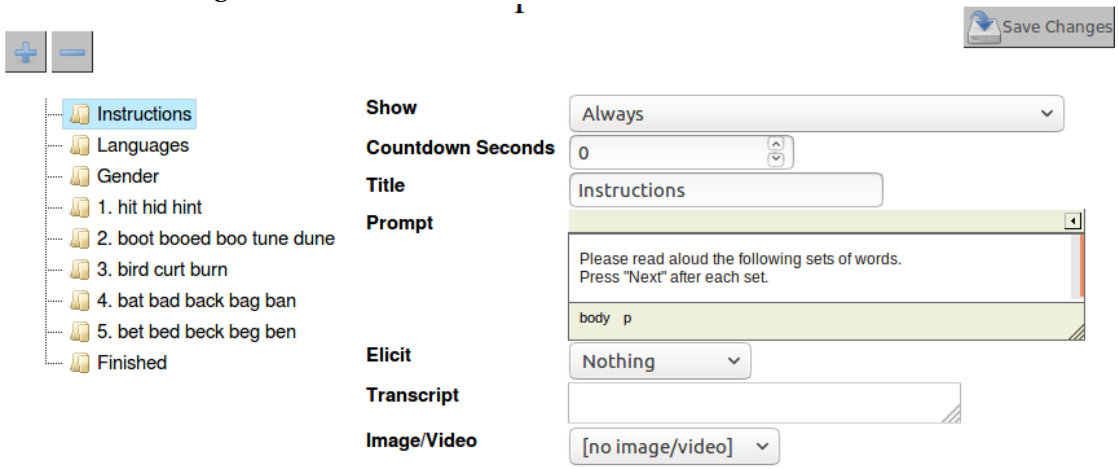
- **Show:** *Always*
- **Countdown Seconds:** 0
- **Title:** (leave this box blank)
- **Prompt:** Please read the following aloud:
- **Elicit:** *Audio*
- **Transcript:** `1. hit hid hint`
- **Max Seconds:** 30
- **Next Button:** *Shown*
- **Rerecord Button:** *Hidden*
- **Image/Video:** *no image/video*

(62) Press the button to add a new step
(63) Fill in the same details as the previous step, except:
Transcript: `2. boot booed boo tune dune`
(64) Add a new step for **Transcript:** `3. bird curt burn`
(65) Add a new step for **Transcript:** `4. bat bad back bag ban`
(66) Add a new step for **Transcript:** `5. bet bed beck beg ben`
(67) Add one last step, with the following details:

- **Show:** *Always*
- **Countdown Seconds:** 0
- **Title:** Finished
- **Prompt:** Thanks for your participation!
- **Elicit:** *Nothing*
- **Transcript:** (leave this box blank)
- **Image/Video:** *no image/video*

This last step is what is displayed to the participant when they've finished all the steps.

(68) Press *Save Changes*



Your task is almost ready. We just need to define which options for *gender* they can see.

(69) Select the *elicitation tasks* option on the menu.
(70) Press *Participant Attributes*.
(71) Select *Options*
This displays a list of the gender options that are visible to the participant. As you can see it's currently empty. In this case, we want to display all options for them to select.
(72) Press *Add All*
You will see that all the options (`M`, `F`, and '(not specified)') have been added to the list. If you wanted to, you could edit the "description" of the individual items (e.g. translate them to another language if your participants don't speech English), or delete options you don't want them to be able to select.
(73) Press the *Delete* button next to the '(not specified)' option, and click *OK* to confirm.
(74) Press *Save Changes*.

Your task is now fully defined and ready to go.

Now you're going to run through the elicitation task yourself …

(75) Select the *elicitation tasks* option on the menu.
(76) Press the *Elicitation Task* button on the bottom right.
You should see a page that displays the task's 'preamble' that you defined earlier.
(77) Click *Next*.
You should see a page that displays the task's *consent* form that you defined earlier, with a box to enter your name in order to 'sign' the consent.
(78) Enter your name and click *Next*.
You will be given the chance to save your copy of the consent form.
(79) Save the consent form and open it to check the contents.

(80) Close the consent form to return to the task.
You will be asked for the demographic details you defined earlier.

(81) Fill in your languages and press *Next*.
(82) Fill in your gender and press *Next*.

You should see a page with some text about enabling your microphone.

If you don't, and instead see a message about your browser not being supported, this means that your web browser doesn't support recording sound. In this case, copy the address of the page at the top, and paste it into another browser (e.g. Google Chrome or Mozilla Firefox).

Once you've enabled your browser for access to your microphone, the task steps will begin, and you should follow the instructions, reading the prompts aloud and clicking *Next* after each group of words.

Each time somebody performs the task, they're assigned a unique Participant ID, which is linked to their demographic data and the recordings.

(83) Press the *Back* button on your browser to return to the *define elicitation tasks* page in LaBB-CAT.
(84) Press the *participants* option on the menu.

> **Tip**
>
> LaBB-CAT remembers the last filters you used, so you may need to clear any filters you had previously applied.
> For example, if the last time you accessed the *participants* page, you selected the `F` gender option to show only female participants, that filter may still be active.
> You can press the ◪ button at the top, to the right of the page filters, to clear all filters.

(85) Under the *Corpus* heading, select the *CC* option.
You will see one participant; the one you just created by doing the task.
(86) Press the participant ID to open their attributes page, and check that the demographic information you entered has been saved.
(87) Click the participant ID to open their attributes page, and check You will see that the participant has five transcripts, one for each of the task steps where audio was recorded.
(88) Press the *Transcripts* link at the bottom to list the transcripts.
(89) Open the first transcript.
You will see that the transcript starts with a comment, which is the prompt text you were shown during the step, and that the transcript contains one utterance.
(90) Play the audio to ensure it was recorded correctly.
(If the last transcript you looked at had video, you may need to tick the checkbox next to the "wav" option in the top right corner, in order to select audio for playback.)

(91) On the right hand edge of the page, about halfway down, there is a green arrow icon .
Press it.
You will see that this opens the next transcript in the 'episode' that you just recorded- i.e. the next set of words you read out. The green arrows on the left and the right of the screen allow you to navigate between the different transcripts in the same recording episode.

Although these 'task step' transcripts are very short, they behave the same as any other transcript; they can be exported, annotated, searched, etc.

You now have a small database with a number of speakers in it, so we can start creating some annotations and doing some searches ...