

## Exercise 2      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.

---

### 1 Manual Upload

1. In LaBB-CAT, click the *upload* option in the menu.
2. Click the first option, ‘upload new transcripts’.
3. Now you need to specify how many transcripts we are going to upload in the series. A series is a set of transcripts that belong together because they were recorded during the same session. In our case, each recording session has only one recording, so enter 1.
4. Click *Upload*.
5. Next to *Transcript:* click *Choose File* and select the file in the “QuakeStories” folder called: “BR178LK\_MargaretSpencer.eaf”.
6. Next to *Media:* click *Choose File*.
7. Each transcript has an audio file and a video file, and you want to upload both. Click the file called “BR178LK\_MargaretSpencer.mp4”, then hold down the < Shift > key on your keyboard and click the file called “BR178LK\_MargaretSpencer.wav”. Then click *Open*.
8. Ensure the *Corpus:* option is ‘QB’
9. Ensure the *Type:* option is ‘interview’
10. Click *Upload*  
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.
11. LaBB-CAT has analysed the structure of the ELAN transcript and pre-selected some default options for layer mappings. For the demo data, these defaults are correct, so you needn’t change anything. Click *Set Mappings* to continue.  
This will display a page listing all the speakers in the transcript, so you can select which one is the ‘main participant’, which is the speaker selected by default for searches and other processing.
12. Ensure that “BR178LK\_MargaretSpencer” is ticked, and the interviewer is not ticked, and click *Set Main Participants*.  
This will display a page with the name of transcript you uploaded, with an *edit meta-data* link, and a progress bar (which may have already finished).
13. Click *edit meta data*  
This will display the attributes for the transcript.

14. Check that you remembered to set *Type:* to ‘interview’. If not, you can fix that on this page, and click *Save Changes*.
15. Below the transcript attributes is a list of participants.  
Click the *Attributes* button for “BR178LK\_MargaretSpencer”.  
This will display the participant attributes we defined in an earlier exercise.
16. 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 click *Save*.
17. Below the participant attributes, there is a list of transcripts that the speaker appears in (in this case, only one). Each has various buttons; to find out what each button does, hover your mouse over it, and a ‘tip’ will appear that tells you what the button does.  
Click the *Transcript - html with sound* button for the “BR178LK\_MargaretSpencer.eaf” transcript.
18. You will now see LaBB-CAT’s ‘interactive transcript’ page for the transcript.  
At the top there is a heading, a list of speakers, and then below this, the lines from the transcript, their speakers in the margin. This includes the words the participants utter, and also any noises, comments, and other events that were put in the transcript in ELAN.  
In the top right corner are some playback controls; click the play button. You will see a shaded rectangle following the participant’s speech.  
Try the other controls to see what they do.
19. Now click on any word in the transcript.  
You will see a menu appear, with options for the ‘Utterance’ (the line), and the word.  
Click the play option in the menu to see what it does.
20. Click on the *formats* link under the title.  
You will see a menu, which includes various formats for exporting the transcript.
21. Select ‘Text Only’
22. Click *Convert*
23. Save the resulting file on your desktop, and then open it.  
You will see the transcript in plain-text form.
24. Back in LaBB-CAT, click the browser’s *back* button to return to the transcript.
25. Click the *formats* link, and select the ‘Praat Text Grid’ option.
26. Save the resulting file on your desktop, and then open it with Praat.  
You will see that the TextGrid has various tiers, two for utterances (one for each speaker), and two for individual words (one for each speaker).  
(You will see that each individual word has a ‘default’ alignment – i.e. the words are evenly spread out during the duration of the line they’re in. In a later exercise we will look at ways to make these word alignments actually line up with the words in the audio signal)

## 2 Batch 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 ‘batch upload’ utility.

27. In LaBB-CAT, click the ‘upload’ option on the menu.
28. Click the *upload transcript batch* link.
29. Save and open the file that appears.
30. If you are asked whether to allow Java to run an application called “LaBB-CAT Batch Transcript Uploader”, allow it.
31. This shows a window with a large blank area in the middle with various buttons below it.  
Right-click on the blank area in the middle, and select ‘Add Transcripts’ from the resulting menu.

32. A window will appear with the title “Select transcripts or folders to add”.  
Navigate to the LaBB-CAT Workshop data folder, select the folder called “QuakeStories”, and click *Open*.  
A progress bar may appear as the utility checks through the folder and its subfolders for transcripts. Once it’s finished, the previously blank area will contain a list of transcripts. Each transcript should have a value filled in for each column – Corpus, Series, Transcript, and Media.
33. Most of the transcripts are monologues, so set *Transcript Type:* to ‘monologue’ on the top right.
34. Click the *Upload Transcripts* button on the bottom left.  
You will see that in the “Progress” column, the text changes to “Transferring” for the first transcript. Then this changes into a progress bar, and once it’s complete, the next transcript changes to “Transferring”, and so on.  
There is an overall progress bar at the bottom. Once it gets to the end and says “Finished”, all the transcripts have been uploaded. One of the transcripts will fail to upload: “BR178LK\_MargaretSpencer.eaf”. That’s because we already uploaded it manually.
35.  While the files are uploading, click the online help link next to the upload transcript batch link you clicked above and read the conditions that must be met in order to use the batch uploader.  
You may also be interested in finding out about the other functions of this utility.
36. Once the uploader is finished, you can close the batch uploader window.
37. To verify that all the transcripts are there, click the *transcripts* option on the menu in LaBB-CAT.  
You should see a list of ten transcripts, and at the bottom, various links to other pages. The text of each page link reflects the name of the first transcript on that page.
38. Use the “search”  box to find UC013AM\_Dom.eaf (You can type just part of the name if you like, and hit < Enter >)
39. Click the *Attributes* button for “UC013AM\_Dom.eaf”.
40. Change *type:* to ‘interview’ and click *Save*.
41. Similarly, the following transcripts are interviews, so change their type accordingly
  - UC215YW\_DanielaMaoate-Cox.eaf
  - UC226AD.eaf

If you search for a transcript and no results are returned, double check the *Corpus:* setting, and change it to ‘[all]’ if necessary.

### 3 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.

42. In LaBB-CAT, click the ‘upload’ option on the menu.
43. Select the *upload participant data* option.
44. Click *Choose File*, and select the file in the LaBB-CAT Exercises data folder called “participants.csv”.
45. Click *Upload*
46. 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.
47. 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 this column’ 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 *age\_category*: → the ‘Age’ LaBB-CAT attribute
- The CSV column *ethnicity*: → the ‘Ethnicity’ LaBB-CAT attribute
- The CSV column *grew\_up*: → the ‘grew\_up’ LaBB-CAT attribute
- The CSV column *grew\_up\_region*: → the ‘grew\_up\_region’ LaBB-CAT attribute
- The CSV column *grew\_up\_town*: → the ‘grew\_up\_town’ LaBB-CAT attribute
- The CSV column *languages\_spoken*: → the ‘languages\_spoken’ LaBB-CAT attribute

48. Click *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:

49. Click the *participants* option on the menu.

You will see a list of ten speakers, and page links at the bottom.

50. Pick a speaker (e.g. QB702\_AnnaSoboleva) and click their *Attributes* button.

You will see the participant attributes page with their details filled in (e.g. QB702\_AnnaSoboleva is a female English/Russian speaker between 18 and 25 years old).

## 4 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 . . .

51. In LaBB-CAT, click the ‘corpora’ option on the menu.

52. Add a corpus called “CC” with a description “Canterbury Corpus”.

53. Click the ‘transcript types’ option on the menu.

54. 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.

55. Click the ‘elicitation tasks’ option on the menu. The page you see is a list of elicitation tasks defined, which is currently empty.

56. 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, will be made into a PDF file which is uploaded with their recordings, and is made available for them to download.

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.

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

57. Click *New* to add the task.

58. Click *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.

59. 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.

60. Close the online help page to return to the “define elicitation steps” page.

61. Fill in the following details:

- *Countdown Seconds*: → 0
- *Title*: → **Instructions**
- *Prompt*: → **Please read aloud the following sets of words. Press "Next" after each set.**
- *Transcript*: → (leave this box blank)
- *Recording*: → ‘Don’t record audio’
- *Image/Video*: → ‘no image/video’

62. Click the  button to add a new step

63. Fill in the following details:

- *Countdown Seconds*: → 5
- *Title*: → (leave this box blank)
- *Prompt*: → **Please read the following aloud:**
- *Transcript*: → 1. **hit hid hint**
- *Recording*: → ‘Record audio’
- *Max Seconds*: → 30
- *Image/Video*: → ‘no image/video’

64. Click the  button to add a new step

65. Fill in the same details as the previous step, except:

*Transcript*: → 2. **boot booed boo tune dune**

66. Add a new step for *Transcript*: → 3. **bird curt burn**

67. Add a new step for *Transcript*: → 4. **bat bad back bag ban**

68. Add a new step for *Transcript*: → 5. **bet bed beck beg ben**

69. Add one last step, with the following details:

- *Countdown Seconds*: → 0
- *Title*: → **Finished**
- *Prompt*: → **Thanks for your participation!**
- *Transcript*: → (leave this box blank)
- *Recording*: → ‘Don’t record audio’
- *Image/Video*: → ‘no image/video’

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

70. Click *Save Changes*

The screenshot shows the configuration interface for a task named 'Instructions'. On the left, a tree view lists the task steps: '1. hit hid', '2. boot bo', '3. bird cu', '4. bat bad', '5. bet bed', and 'Finished'. The main configuration area on the right includes:

- Sub-steps:** A dropdown menu set to 'All steps in this order'.
- Countdown Seconds:** A text input field containing '0'.
- Title:** A text input field containing 'Instructions'.
- Prompt:** A large text area containing the text: 'Please read aloud the following sets of words. Press "Next" after each set.'
- Transcript:** An empty text area.
- Recording:** A dropdown menu set to 'Don't Record Audio'.
- Image/Video:** A dropdown menu set to '[no image/video]'.

You have now defined the steps in the task. Next we'll define what demographic information we will ask each participant before they start. In this case, we will ask for their gender and what languages they speak.

71. Click the 'elicitation tasks' option on the menu.

72. Click *Participant Attributes*. This displays a list of predefined participant attributes that are assigned to the task, although the list is currently empty.

73. Under *attribute:* select 'languages\_spoken'

74. Under *label:* enter Languages

75. Under *description:* enter What languages do you speak?

76. Under *order:* enter 1

77. Click *New*

This adds the "languages\_spoken" attribute to the list.

78. Add another attribute, by filling in the blank row at the bottom:

- *attribute:* → 'gender'
- *label:* → Gender
- *description:* → Are you male or female?
- *order:* → 2

79. Click *New*

Gender has a set of options defined for it, and we need to specify which of the options we will allow them to use, and how those options will be labelled ...

80. Click the *Options* button on the new "gender" row

81. Under *value:* select 'F'

82. Under *descript:* enter Female

83. Click *New*

84. Now add another option by filling in the blank row at the bottom:

- *value:* → 'M'
- *descript:* → Male

85. Click *New*

Your task is now fully defined and ready to go.

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

86. Click the 'elicitation tasks' option on the menu.
87. Click the *Elicitation Task* button on the bottom right.  
You should see a page that displays the task's 'preamble' that you defined earlier.
88. 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.
89. Enter your name and click *Next*.  
You will be given the chance to save your copy of the consent form.
90. Save the consent form and open it to check the contents.
91. Close the consent form to return to the task.  
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).
92. Click *Next* and follow the instructions.

Once you've enabled your browser for access to your microphone, you will be asked for the demographic details you defined earlier.

After you enter these, the task steps will begin, and you should follow the instructions, reading the prompts aloud and clicking *Next* after each group of words.

When you get to the end of the task, you will see a "Participant ID" displayed; each time somebody performs the task, they're assigned a unique ID, which is linked to their demographic data and the recordings.

93. Click the *Back* button on your browser to return to the "define elicitation tasks" page in LaBB-CAT.
94. Click the 'participants' option on the menu.
95. Selected the 'CC' corpus from the list.  
You will see one participant; the one you just created by doing the task.
96. Click the *Attributes* button and check that the demographic information you entered has been saved.  
You will see that the participant has five transcripts, one for each of the task steps where audio was recorded.
97. 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.
98. Play the audio to ensure it was recorded correctly.

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 . . .