2 - Upload Data

Uploading Transcripts and Recordings

Transcripts can be uploaded one at a time, which allows the manual selection of specific options during upload, like corpus, transript type, main participant, etc.

However, if you already have a collection of transcripts and media files (which we have for these exercises – download QuakeStories.zip to get the workshop data) – and they are systematically organised (which they are), you may be able to save some manual uploading work by uploading them using the 'batch upload' utility, which automates some of the decision-making.

1. In LaBB-CAT, select the *transcripts* option in the menu.
2. Press the *Upload Transcript Batch* icon at the bottom.
   This shows a window with a large blank area in the middle with various buttons above it.
3. Open Windows Explorer or Finder, and navigate to the LaBB-CAT Workshop data folder.
4. Drag the folder called "QuakeStories", and drop it on to LaBB-CAT, on to the blank area below the buttons.
   The previously blank area will contain a list of transcripts. Each transcript should have a value filled in for each column - *Transcript*, *Media*, *Corpus*, and *Episode*.
5. Most of the transcripts are monologues, so set *Type* to *monologue* on the top left.
6. 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.
7. While the files are uploading, click the online help link next to the upload transcript batch link you used above and read the conditions that must be met in order to use the batch uploader.
8. Once the uploader is finished, 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.
9. Use the "Transcript" box to find *UC013AM_Dom.eaf*
   (You can type just part of the name if you like)
10. Select the *Attributes* icon for *UC013AM_Dom.eaf*
    (the one with the spanner/wrench icon on the right).
11. Change *type* to *interview* and press *Save*.
12. Similarly, the following transcripts are interviews, so change their type accordingly

    - UC215YW_DanielaMaoate-Cox.eaf
    - UC226AD.eaf

13. The heading at the top of the transcript attributes page, which is the name of the transcript, is a link. Click the link.

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 annotations that were put in the transcript in ELAN.

14. In the top right corner are some playback controls; press the play button. You will see a shaded rectangle following the participant's speech.
15. Try the other controls to see what they do.
16. Now click on any word in the transcript.
    You will see a menu appear, with options for the 'Utterance' (the line), and the word.
    Press the play option in the menu to see what it does.
17. Click on the *formats* link under the title.
    You will see a menu, which includes various formats for exporting the transcript.
18. Select *Plain Text Document*
19. Save the resulting file on your desktop, and then open it.
    You will see the transcript in plain-text form.
20. Click the *formats* link, and select the *Praat Text Grid* option.
21. 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. It is possible to make these word alignments actually line up with the words in the audio signal, using forced aligment, which is the subject of another tutorial.)

You can also open individual utterances in Praat directly from the transcript page, if you have Praat installed. But first, the LaBB-CAT/Praat integration has to be set up; this only has to be done once:

21. On the top-right of the page, above the playback controls, there's a Praat icon - click it.
22. Follow the instructions that appear (these vary depending on what web browser you use).

> **Tip**
>
> You may need to grant a browser extension permission to install, and it's possible you will need a connection to the internet in order to download this extension.

You also may be asked where Praat is installed; Navigate to the location where Praat is installed, and double-click the “Praat.exe” file (on some systems the file may simply be called ”Praat”). The Praat program may open, and then immediately close, as LaBB-CAT tests it can communicate with Praat.

 There are illustrated instructions for setting up Praat integration for each web browser in the online help for the transcript page; check there if you run into problems.

Now Praat integration has been set up, and you should be able to access Praat options in the transcript page from now on…

23. Click on a line in the transcript, and select the *Open Text Grid in Praat* option on the menu.

Praat should open, and show you a spectrogram of the line’s audio, with a TextGrid below that includes a tier for the utterance, and another tier for individual word alignments. You could manually align them here, but it’s much more efficient to use HTK to force-align the utterances. Forced alignment is the subject of another tutorial…

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.

1. In LaBB-CAT, select the *participant* option in the menu.
2. Press the *Upload Participant Data* icon at the bottom.
3. Press *Choose File*, and select the file in the LaBB-CAT Exercises data folder called “participants.csv”.
4. Press *Upload*
   You will now see a list of the columns from the spreadsheet.

5. 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.
6. Below that is listed each column from the spreadsheet, with an arrow pointing to a drop-down 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. Select the options as follows 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 *create a new attribute called* option, and set the **Label** to "Age"
- The CSV column **ethnicity:** → the *create a new attribute called* option, and set the **Label** to "Ethnicity"
- The CSV column **grewUp:** → the *create a new attribute called* option, and set the **Label** to "Country"
- The CSV column **grewUpRegion:** → the *create a new attribute called* option, and set the **Label** to "Region"
- The CSV column **grewUpTown:** → the *create a new attribute called* option, and set the **Label** to "Town"
- The CSV column **languagesSpoken:** → the *create a new attribute called* option, and set the **Label** to "Languages"

7. 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:

8. 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.
9. Pick a speaker (e.g. *QB702_AnnaSoboleva*) and click their name.
   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).

By default, the new attributes are not flagged as searchable, so we will make a few of them searchable now.

10. Select the *participant attributes* link on the menu.
    This will display a list of the participant meta-data fields.
11. Ensure that *Searchability* is set to *Searchable* for the following attributes:
    - *gender*
    - *ageCategory*
    - *languagesSpoken*
12. Press the *Save* button at the bottom of the list.
13. Select *participants* on the top menu.
    You will see that the searchable attributes are now listed with the participants.

    You can filter the list using the attribute headers at the top of the list.

14. Under *Gender* select *F*
    Now the list only shows female participants.

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