2 - Search

In addition to storing recordings and orthographic transcripts, the data can also be annotated in various ways with different information. Each type of annotation is stored on its own 'layer', so you can display and search on the basis of different aspects of the transcripts, including:

- frequency
- lemma
- part of speech
- pronunciation
- speech rate
- pause duration
- …and more.

Annotations can be made manually, and LaBB-CAT includes modules (called 'Layer Managers') for doing certain annotations automatically.

Various automatically generated annotation layers have been configured in the demo instance of LaBB-CAT, and we will start to explore some of them in this worksheet.

Layered Search Matrix

Layered search is usually 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.

1. Firstly, return to LaBB-CAT's home page by clicking the *Home* link on the menu, and then click the *Layered Search* icon.
   You will see a page called "Participants".
2. Select 'M' in the *Gender* box.
   You will see a list of the male participants only.
   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 un-ticked.)
3. Press the *Layered Search* button at the top of the list.
   You will see a page that lists the speakers at the top, a number of tickable annotation layers in the middle, 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.)

4. In the box labelled "orthography" type the regular expression `th[aeiou].+`
   As you saw earlier, `[aeiou]` means 'any vowel', and a full-stop/period means 'any character'
   The plus-sign means 'one or more of the previous thing', so `.+` means 'at least one character'.
5. Now press the *Search* button at the bottom (or hit `Enter`).
   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 that the results include words like "that", "there", "then", etc. - i.e. words that start with "th", followed by a vowel, followed by at least one more letter.

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

As we previously saw with the 'easy search', each match is highlighted and shown within a few words context. However, this results page has a few more options available.

6. In the *Context* drop-down box at the top, select the *5 words* option, to show more context in the llist of results.
7. Each result line has a ticked checkbox next to it. At the bottom of the list, you'll see that there are various buttons, which perform operations on the ticked results, including *CSV Export*, *Utterance Export*, and *Audio Export*.
8. Untick the "*Select all result*" 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.

9. Click the *Audio Export* button.
10. Save and open the resulting zip file.
    You'll see that extracted wav files are systematically named to include:

    - the name of the transcript
    - the start and end time of the extracted utterance

11. If you also have Praat installed on your computer, go back to the results page and click *Utterance Export* button. Save and open the resulting zip file.

You'll see that the TextGrid names match the audio file names in the previous zip file. If you open a TextGrid in Praat, you'll see it includes a tier for the whole utterance transcript, a tier with an interval for each word, and a *target…* tier which tags the word that matched the regular expression you searched for.

12. Back on the results page, click the *CSV Export* button.
13. 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.

> **Note**
>
> If you're using Microsoft Excel and you find it doesn't open all the columns correctly:
>
> 1. Create a new workbook in Excel.
> 2. Click the 'Data' tab.
> 3. On the "Get External Data" ribbon click 'From Text'.
> 4. Select the CSV file you downloaded.
> 5. Select 'Delimited' and click *Next*.
> 6. Ensure 'Comma' is the only delimiter ticked and click *Next*.
> 7. Click *Finish* and then *OK*.

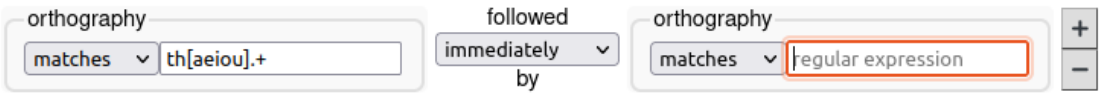
You will see a spreadsheet with one line per selected result, and various columns containing information about the speaker, the corpus, the match line and word, and a URL to the interactive transcript for the match.

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…

14. Close the CSV file, and got back to the results page.

Up until now, we've only been matching against one word at a time. Now we're going to include patterns for a chain of words. Unlike the simple search, adding a space in the regular won't work, because each column in the search matrix only matches a single word. To match a chain of two words, we need to have two columns in the search matrix.

15. On the search page, next to the *orthography* box where you entered the regular expression, there's a button for adding a column to the matrix. Click it.



    Now you will see that our search matrix is one layer high by two words wide.
16. Change the entries on the *orthography* layer so that it will match the word "the" followed immediately by a word that starts with a vowel, and click *Search*.

Check the search results are giving you what you expected. You may note that some of the following words start with a vowel in the spelling, even though they are not *pronounced* with a vowel sound. We will see how to search on the basis of pronunciation in another worksheet.

17. Now search for "the" followed, within two words, by a word that starts with a vowel.

> **Tip**
>
> If in doubt about a search option, try the online help page.

Searching Other Layers

So far we have only searched the *orthography* layer - i.e. the ordinary spellings of words. But LaBB-CAT has been configured to generate a number of other annotation layers.

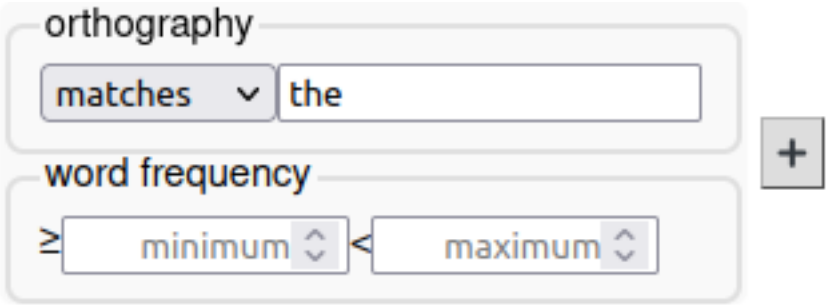
Let's say we're interested in how rare or common words are in our data.

LaBB-CAT's 'Frequency Layer Manager' is a module that counts up the number of times each word type appears in the database. It generates a frequency list, and also annotates each word token with its frequency.

We'll now search for tokens of words that appear only once in the database.

The annotation layers are grouped into a number of 'projects' to avoid clutter. We will initially be interested in the layers related to frequency.

1. On the search page, in the box labelled *Tick layers to include*, there's a *Projects* column. Tick the *frequency'* project.
   Some addition layers will appear in the layer list on the right.
2. Tick the *word frequency* layer.
3. Set the word matrix to be 1 word wide again by clicking the button to the right.
   You will see that the search matrix now has two layers in it.



4. Unlike the *orthography* layer, which has one box for a regular expression, the *word frequency* layer has two boxes, marked " " and "<". This is because the annotation values are numbers.
   We want all the words that appeared only once in the database. Enter a number or numbers in the appropriate box (you can leave either box blank) and click *Search*.

> **!** Important
>
> Ensure the orthography box is empty, otherwise it will be trying to find in instances of the word "the" that appear only once in the corpus; there are lots of instances of the word "the", so the search will return no results, as the frequency is greater than 1.

5. Click on the first result in the list.

Layered Transcript

This displays the 'layered transcript' page for the recording. This is similar to the previous 'easy' transcript page, but has a number of extra options and functions.

The most obvious difference is that each word token has a number above it. This is the frequency of that word, which is displayed because the *word frequency* layer is selected; there's a list of layers at the top of the transcript, and you can see that both *word frequency* and *word* are ticked.

1. Untick the *word frequency* layer.
   After a short delay, the transcript will be displayed again, with only the transcript text visible.

The transcript also includes any noises (e.g. "tuts"), comments, and other events that were put in the transcript in ELAN.

2. The video is the top right corner as before; click the play button. Again you will see a shaded rectangle following the participant's speech.
3. Try clicking the magnifying-glass icon below the video, to see what it does.
4. Now click on any word in the transcript.
   You will see a menu appear.
5. Click the play option in the menu to see what it does.
6. Click on the *formats* link under the title.
   You will see a menu, which includes various formats for exporting the transcript.
7. Select *Plain Text Document*
8. Save the resulting file on your desktop, and then open it.
   You will see the transcript in plain-text form.

9. 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).

   (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)

Using frequencies of full wordforms can be useful, but in some circumstances it may be more informative to group together different forms of the same word; e.g. treat "damage", "damaged" and "damaging" as variants of the same thing for the purposes of frequency-counting.

We'll see a way to do that in the next worksheet.

In this worksheet you have seen that:

- Annotations can be automatically added to transcripts in layers, using Layer Managers.
- The Frequency Layer Manager can tag words with their frequencies, and maintains a frequency list.
- Annotation layers can be searched using the search matrix, using numeric value or regular expressions.
- Layers can be optionally displayed in transcripts.