

Exercise 2 Layered 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.

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. Click the *search* option on the menu and click *Search Everyone*
2. Under the “projects” heading, tick the ‘frequency’ project.
Some additional layers will appear in the layer list on the right.
3. Tick the *word frequency* layer.
4. Press the *Set Search Matrix* button below.

You will see that the search matrix now has two layers in it.

anything ▾ ¹² word frequency ≥ < followed by ▾
followed by 🌀 orthography ▾ anything ▾

5. 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*.
6. Click on the first result in the list.

This displays the transcript page you’ve already seen. However, now each word token has a number above it.

This is the frequency of that word, which is displayed because the *word frequency* is selected; there’s a list of layers at the top of the transcript, and you can see that both *word frequency* and *transcript* are ticked.

You should see that the highlighted word that matched your search has a “1” above it (which means that word appears only once in this corpus), and that other words in the transcript are tagged with other numbers.

(You will see the transcript also includes any noises (e.g. “tuts”), comments, and other events that were put in the transcript in ELAN; these are annotations on their own layers.)

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

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.
- Annotation layers can be searched using the search matrix, using numeric value or regular expressions.
- Layers can be optionally displayed in transcripts.