

Exercise 3 Searching

Now that you have some transcripts in your database, we'll try out LaBB-CAT's search functions a little. Searching broadly involves the following steps:

1. Selecting participants whose utterances you want to search,
2. Specifying one or more patterns to search for, and
3. Exploring or extracting the search results.

We'll start with a very simple search – all the instances of the word “the” uttered by monolingual English-speaking males.

1. In LaBB-CAT, click on the *search* link on the menu.
This takes you to a page entitled “filter”, where you can list participants and filter them by their attributes. You can see various participant attributes listed across the page.
2. We're interested in male participants, so under the word “Gender”, select ‘Male’ from the list.
The page will then display a list of all the male participants in the database.
3. We want the participants who speak only English, so select ‘English’ under “Languages Spoken”
The page will then display a list of male participants who list only “English” as their language.
4. Click *Layered Search* at the bottom.
You will see the participants you selected listed at the top, next to a list of layers (which we'll ignore for now). Below that, there's a heading “search” with various controls. This is the ‘search matrix’, although it doesn't look much like a matrix yet, because it's only one layer high and one word wide. . .
5. In the box next to the word “orthography” type the word **the**

6. Now click the *Search* button at the bottom.
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.
7. Each match is highlighted and shown within a few words context. Click on the first match.
You will see that the interactive transcript page opens in a new browser tab, with the match at the top, and highlighted. You will also see that all the other matches from the same transcript are also highlighted.
8. We've already seen what can be done in the interactive transcript page, so close the tab to return to the results page.
9. Each result line has a ticked checkbox next to it. Scroll to the bottom of the list.
You'll see that there are three buttons at the bottom, which perform operations on the ticked results – *CSV Export*, *Extract Audio*, and *Convert*.
10. Untick the “[select all *N* results]” 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.
11. Click the *Extract Audio* button.
12. Save and open the resulting zip file.
You'll see that the files are systematically named to include:
 - the result number
 - the name of the transcript
 - the start and end time of the extracted utterance

13. Now go back to the results page and select ‘Praat TextGrid’ from the dropdown list next to the *Convert* button, and then click *Convert*.
14. Save and open the resulting zip file.
You’ll see that the TextGrid names match the audio file names in the previous zip file.
15. Back on the results page, click the *CSV Export* button.

16. 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.

Tip: If you’re using Microsoft Excel and you find it doesn’t open all the columns correctly:

- (a) Create a new workbook in Excel.
- (b) Click the ‘Data’ tab.
- (c) On the “Get External Data” ribbon click ‘From Text’.
- (d) Select the CSV file you downloaded.
- (e) Select ‘Delimited’ and click *Next*.
- (f) Ensure ‘Comma’ is the only delimiter ticked and click *Next*.
- (g) 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...

17. Close the CSV file, and the results page, and go back to the search matrix page.

We’ve seen that you can search for exact word matches, but you can also search for patterns, using ‘regular expressions’. Now we’re going to search for words *beginning with* “the...”

18. Change the *orthography* search text to **the.*** (i.e. after the word “the”, append a full-stop and an asterisk).

The screenshot shows a search interface with a dropdown menu set to 'anything', a search icon, the word 'orthography' in a dropdown, a text input field containing 'the.*', and a 'followed by' dropdown set to 'anything'.

The full-stop means “any character at all”, and the asterisk means “zero or more of the previous thing”, so **.*** means “zero or more characters”.

19. Click *Search*.
You will see that now the search results include the word “the” and also words like “then”, “there”, “they”, etc.
20. Now go back to the search page, and change the asterisk to a plus-sign, which means “one or more of the previous thing”

The screenshot shows the search interface with the text input field now containing 'the+'.

21. Click *Search*.
You will see that now the search results exclude the word “the”, only including words where the initial “the...” is followed by at least one character.
22. Now change your search by replacing the **e** in “the” with **[aeiou]** – so your search pattern will be **th[aeiou].+**
The square-brackets mean “any one of the things inside the brackets”, so **[aeiou]** means “any vowel”.

23. Click *Search*.
You will now see that the results include words like “think”, “that”, “thought”, etc.

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

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...

24. On the search page, underneath the list of layers, there's a box with the number 1 in it. Change the number to 2 and click *Set Search Matrix*.

The screenshot shows a search interface with the following elements:

- A dropdown menu with 'anything' selected, followed by the text 'followed by'.
- A gear icon followed by the layer name 'orthography' and a dropdown arrow.
- A text input field containing the regex 'th[aeiou].+'.
- A dropdown menu with 'immediately' selected, followed by a dropdown arrow.
- An empty text input field.
- The text 'followed by' above the empty field.
- The text 'by' below the empty field.
- A dropdown menu with 'anything' selected.

Now you will see that our search matrix is one layer high by two words wide.

25. 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.
26. Now search for “the” followed, within two words, by a word that starts with a vowel.
27. Dream up some other searches that interest you, and try out other options on the search page. For example you might try finding all words that end in “. . .ing”

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