4. 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, select the *participants* option on the menu.
This takes you to a page listing all participants, where you can filter participants by their attributes. You can see various participant attributes listed across the top of page.

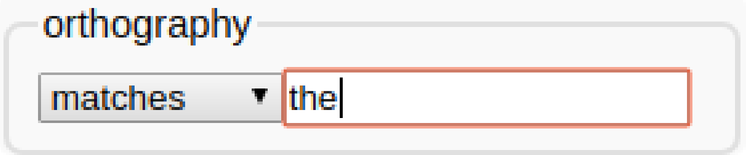
> Tip
>
> LaBB-CAT remembers the last filters you used, so you may need to clear any filters you had previously applied.
> For example, if the last time you accessed the *participants* page, you selected the `CC` corpus to show only *CC* participants, that filter may still be active. To remove that filter, just press `CC` again to de-select that option, or you can press the 🧹 button at the top, to the right of the page filters, to clear all filters.

(2) We're interested in male participants, so under the *Gender* attribute, select *M*.
After a short delay the page will display a list of all the male participants in the database.

(3) We want the participants who speak only English, so enter `English` under *Languages Spoken*
The page will then display a list of male participants who include *English* in their languages. It also includes participants who speak other languages, and we want to eliminate these.
The *Languages* filter box accepts a 'regular expression' for matching patterns, so if we enter `^English$` in the box, only those with English as their sole language will be listed. This is because, in regular expressions, `^` means "the beginning" and `$` means "the end", so `^English$` means, "English at the beginning, and at the end"

(4) Click the *Layered Search* button at the top of the list.
You will see the participants you selected listed at the top, above a list of annotation layers. Below that, there's a "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 under the word "orthography" type the word `the`



orthography

matches ▼ the

(6) Now press the *Search* button at the bottom (or hit `Enter`).
A progress bar will appear, and then shortly after that, a new browser tab 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 with some context (the previous word and the following word in the transcript). The amount of context is controlled by a drop-down list at the top.
Select *5 words* to see more context around each match.

(8) Click on the first match.
You will see that the interactive transcript page opens in a new 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.

(9) We've already seen what can be done in the interactive transcript page, so close the tab to return to the results page.

(10) Each result line has a ticked checkbox next to it. Scroll to the bottom of the list.
You'll see that there are buttons at the bottom, which perform operations on the ticked results, including *CSV Export*, *Utterance Export*, and *Audio Export*.

(11) Un-tick the "Select all 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.

(12) Press the *Audio Export* button.

(13) Save and open the resulting zip file.
You'll see that the files are systematically named to include:

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

(14) Now go back to the results page and tick the *Prefix Names* checkbox.

(15) Press the *Audio Export* button again.

(16) Save and open the resulting zip file.
This time you'll see that the files are also prefixed by the result number.
You may notice that there are more audio files this time; that's because there were multiple results in the same utterance. Previously, only one copy of the utterance was exported, but this time, each match has its own copy of the utterance audio, prefixed by the result number.

(17) Now go back to the results page and un-tick the *Prefix Names* checkbox.

(18) Click the *Utterance Export* button.

(19) Save and open the resulting zip file.
You'll see that the TextGrid names match the audio file names in the first zip file.

(20) Open one of the TextGrids in Praat.
You'll see that the TextGrid includes a tier named *target*… which indicates which token(s) in the *word*… tier matched the search pattern.

(21) Back on the results page, click the *CSV Export* button.

(22) 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.

> ℹ 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*.

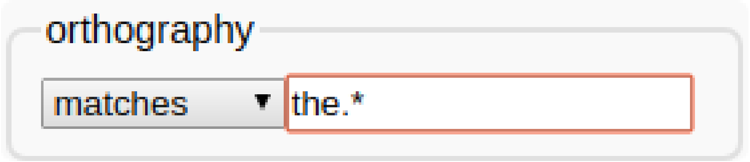
You will see a spreadsheet with one line per selected result, and various columns containing information about the speaker, the corpus, the match line and word, and a URL to the interactive transcript for the match.

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…

(23) 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…"

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



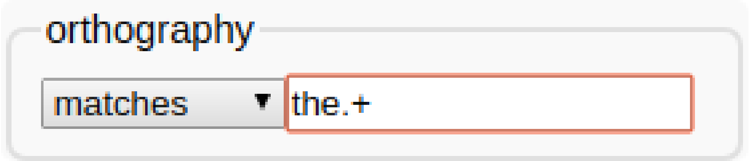
orthography

matches ▼ the.*

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”.

(25) Click *Search*.
You will see that now the search results include the word “the” and also words like “then”, “there”, “they”, etc.

(26) Now go back to the search page, and change the asterisk to a plus-sign, which means “one or more of the previous thing”



orthography

matches ▼ the.+

(27) 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.

(28) 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”

ℹ Note

While you are typing the regular expression, you may notice that the text goes red; this means that what’s currently in the box is not a valid regular expression. That’s fine while you’re still typing, but when you’re ready to search, if the text is red, the search will likely fail. If the regular expression text is red, you can see what the problem with it is by hovering your mouse over the red text; a ‘tip’ will appear showing an error message

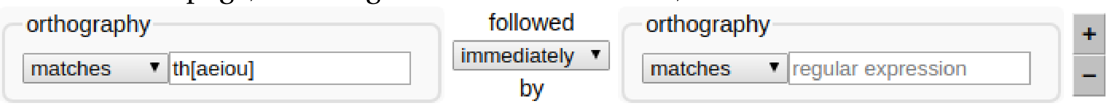
(29) Click *Search*.
You will now see that the results include words like “think”, “that”, “thought”, etc.

Tip

You can get more information about regular expressions by using the online help on the search 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…

(30) On the search page, to the right of the search matrix, there's a + button. Click it.



orthography: matches th[aeiou]

followed immediately by

orthography: matches regular expression

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

(31) 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.

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

(33) Dream up some other searches that interest you, and try out other options on the search page.

Tip

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