This tutorial is a very brief introduction to the LaBB-CAT corpus analysis tool. There are four exercises, in which you:

1. install and configure LaBB-CAT,
2. upload a small corpus of recordings into your database and explore the transcript page,
3. search for some tokens using regular expressions,
4. install and configure a module for automatically annotating tokens with their phonemic transcriptsions, and search for tokens based on pronunciation instead of spelling.

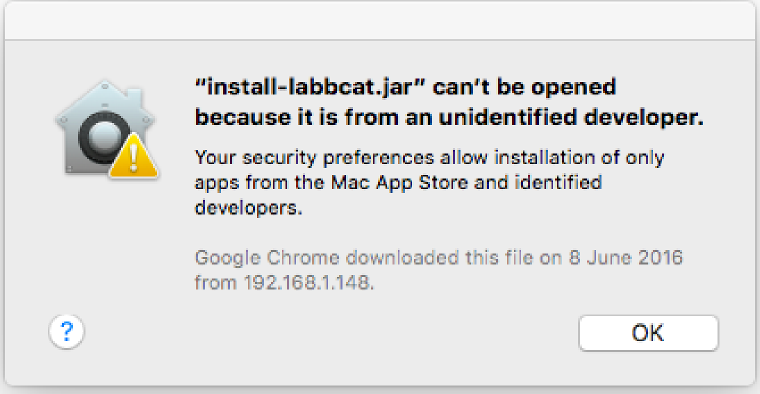
Exercise 1 Setting Up

In this exercise you will:

- Install the LaBB-CAT software
- Define corpora
- Define transcript types

After this you will have an empty LaBB-CAT database set up ready to upload transcripts into.

1. You have a file called *install-labbcat.jar* – double click this file to start the installer.
   If you are using OS X, you may see a message that the file can't be opened:



"install-labbcat.jar" can't be opened because it is from an unidentified developer.

Your security preferences allow installation of only apps from the Mac App Store and identified developers.

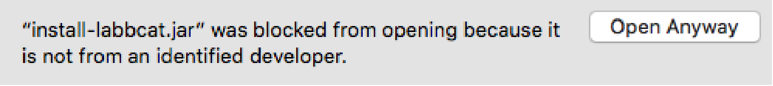
Google Chrome downloaded this file on 8 June 2016 from 192.168.1.148.

If this happens:

(a) Click the Apple icon in the top left corner of the screen.

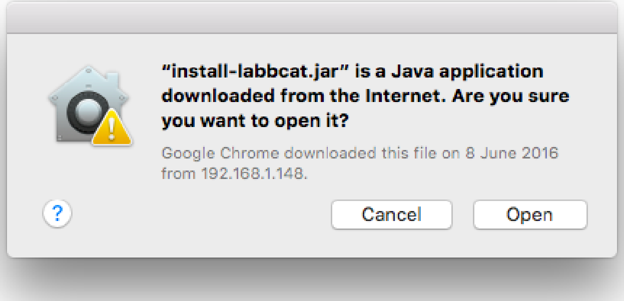
(b) Select *System Preferences*

(c) Click *Security & Privacy*
Near the bottom it says *"install-labbcat.jar" was blocked from opening because it is not from an identified developer.*



"install-labbcat.jar" was blocked from opening because it is not from an identified developer.

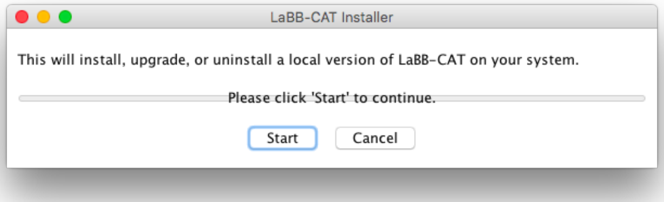
(d) Click *Open Anyway*
You may see another warning about the program being downloaded from the internet



"install-labbcat.jar" is a Java application downloaded from the Internet. Are you sure you want to open it?

Google Chrome downloaded this file on 8 June 2016 from 192.168.1.148.

(e) Click *Open*



This will install, upgrade, or uninstall a local version of LaBB-CAT on your system.

Please click 'Start' to continue.

2. Click *Start* You will see the progress bar move as files are installed. Once this is finished, you'll see a message saying "Installation complete."

3. Click *Finished* to close the installer

The software is now installed. LaBB-CAT is a browser-based system, which means that it works as a mini web server on your computer, and you need to access it using your web browser.

Each time you want to use LaBB-CAT, you must start it up, and which you've finished, you close it down again.

To start LaBB-CAT, click the LaBB-CAT icon in your applications area.

- On Windows, open the *Start* menu and type LaBB-CAT.
- On OS X you will find LaBB-CAT in your *Applications* folder.

A window called "LaBB-CAT Server" will open, and after a short delay, your default web browser will open on a page called "LaBB-CAT".

Now that the software is installed, we will set up a basic structure for receiving data:

4. Start LaBB-CAT as described above.
   This will open your web browser on LaBB-CAT's start page.

5. The very first time this opens, you will see the software's licence. Click *I Agree* to access the start page.

6. The start page has a link on it called "Where do I start?" – you may like to click on this link and read the first section, which explains a little about how to navigate around LaBB-CAT and where to find online help and hints.

7. Click back on the start page of LaBB-CAT (the page with the "Where do I start?" link).

Now we will set up some corpus names. . .

8. On the menu at the top, click the *corpora* link.
   This page shows a list of current corpora, which only contains one corpus, called 'corpus'.

9. Underneath the 'corpus' corpus, there's a form that you can fill in to add a new corpus. Fill in the following information:

   - *name:* `QB`

- *language:* (leave this as 'English')
- *description:* `Quakebox recordings`

10. Click the *New* button to add the 'QB' corpus.
    You should see a message at the top of the page saying "Record created" and now the 'QB' corpus is in the list, under the 'corpus' corpus.

11. Add another corpus called `UC` with the description `Campus recordings`.

12. We won't actually be using the corpus called 'corpus', so we want to delete it. To do this, click the *Delete* button below and to the right of the 'corpus' corpus in the list.

13. You will be asked "Are you sure you want to delete this record?" You are sure, so click *OK*.
    This will 'cross out' the corpus in the list.

14. Now click the *Save Changes* button that has appeared in the top right corner of the page.
    The row will be deleted from the list.

Now you have some corpora set up with the names you've provided.

The data we are using is a collection of stories about peoples' experiences during the devastating earthquakes that hit the Canterbury region of New Zealand in 2010 and 2011. Some recordings are interviews, where an interviewer asks the participant questions, and others are monologues. Now we're going to set up these two transcript types ...

15. Click on the *transcript types* menu option.
    You will see a list of transcript types, although there's currently only one type in the list, called 'interview'.

16. Underneath this, fill in the empty box with the word: `monologue`

17. Click the *New* button.
    You will see a message at the top saying "Record created", and you will notice that now the list has two transcript types, 'interview' and 'monologue'.

Now you have an empty database for which you've:

- created two corpora, QB and UC, and
- created a new transcript type, so that we can have monologues as well as interviews.

Exercise 2 Uploading Data

In this exercise you will:

1. Upload many transcripts at once using the batch uploader
2. Import participant data from a CSV file

After this you will have a small corpus imported into LaBB-CAT.

Transcripts can be uploaded manually, one at a time, using the *upload new transcripts* option in the *upload* menu.

However, if you already have a collection of transcripts and media files (which we have for these exercises), and they are systematically organized (which they are), you can save some manual uploading work by uploading them using the 'batch upload' utility.

3. In LaBB-CAT, click the 'upload' option on the menu.

4. Click the *upload transcript batch* link.

5. Save and open the file that appears.

6. If you are asked whether to allow Java to run an application called "LaBB-CAT Batch Transcript Uploader", allow it.

7. This shows a window with a large blank area in the middle with various buttons below it.
   Right-click on the blank area in the middle, and select 'Add Transcripts' from the resulting menu.

8. A window will appear with the title “Select transcripts or folders to add”.
   Navigate to the LaBB-CAT Workshop data folder, select the folder called “QuakeStories”, and click *Open*.
   A progress bar may appear as the utility checks through the folder and its subfolders for transcripts. Once it’s finished, the previously blank area will contain a list of transcripts. Each transcript should have a value filled in for each column – Corpus, Series, Transcript, and Media.

9. Most of the transcripts are monologues, so set *Transcript Type:* to ‘monologue’ on the top right.

10. Click the *Upload Transcripts* button on the bottom left.
    You will see that in the “Progress” column, the text changes to “Transferring” for the first transcript. Then this changes into a progress bar, and once it’s complete, the next transcript changes to “Transferring”, and so on.
    There is an overall progress bar at the bottom. Once it gets to the end and says “Finished”, all the transcripts have been uploaded.

11. ❓While the files are uploading, click the online help link next to the upload transcript batch link you clicked above and read the conditions that must be met in order to use the batch uploader.
    You may also be interested in finding out about the other functions of this utility.

12. Once the uploader is finished, you can close the batch uploader window.

13. To verify that all the transcripts are there, click the *transcripts* option on the menu in LaBB-CAT.
    You should see a list of ten transcripts, and at the bottom, various links to other pages. The text of each page link reflects the name of the first transcript on that page.

14. Use the “search” 🔍 box to find `UC215YW_DanielaMaoate-Cox.eaf` (You can type just part of the name if you like, and hit < `Enter` >)

15. Click the *Attributes* button for “BR178LK_MargaretSpencer.eaf”.
    This will display the attributes for the transcript.

16. Change *type:* to ‘interview’ and click *Save*.

17. Similarly, the following transcripts are interviews, so change their type accordingly

    - UC013AM_Dom.eaf
    - BR178LK_MargaretSpencer.eaf

    If you search for a transcript and no results are returned, double check the *Corpus:* filter, and change it to ‘[all]’ if necessary.

18. Below the transcript attributes is a row of buttons, click the *Transcript* button.

19. 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.
    In the top right corner are some playback controls; click the play button. You will see a shaded rectangle following the participant’s speech.
    Try the other controls to see what they do.

20. Now click on any word in the transcript.
    You will see a menu appear, with options for the ‘Utterance’ (the line), and the word.
    Click the play option in the menu to see what it does.

21. Click on the *formats* link under the title.
    You will see a menu, which includes various formats for exporting the transcript.

22. Select ‘Text Only’

23. Click *Convert*

24. Save the resulting file on your desktop, and then open it.
    You will see the transcript in plain-text form.

25. Back in LaBB-CAT, click the browser’s *back* button to return to the transcript.

26. Click the *formats* link, and select the ‘Praat Text Grid’ option.

27. 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:

28. On the top-right of the page, above the playback controls, there's a Praat icon – click it.

29. Follow the instructions that appear (these vary depending on what web browser you use).
    You may be asked whether to allow the "LaBB-CAT Integration Applet" to run. If you tick the "Do not show this again" option, then this message will not appear every time you open a transcript.
    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 integeration 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...

30. Click on a line in the transcript, and select the 'Open Text Grid in Praat' option on the menu.
    You may be asked you if want to allow access to the "LaBB-CAT Integration Applet" - if so, tick "Do not show this again", and click *Allow*.
    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...

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.

31. In LaBB-CAT, click the 'upload' option on the menu.

32. Select the *upload participant data* option.

33. Click *Choose File*, and select the file in the LaBB-CAT Exercises data folder called "participants.csv".

34. Click *Upload*

35. You will now see a list of the columns from the spreadsheet.
    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.

36. Below that is listed each column from the spreadsheet, with an arrow pointing to a dropdown box. The box contains various options, including each of the participant attributes set up in LaBB-CAT, an 'ignore this column' option, and 'create a new attribute' option.
    Ensure that the columns are set up as follows:

    - *name:* → 'ignore' because it's the *Participant Identity Column:* identified above
    - *gender:* → the 'Gender' LaBB-CAT attribute
    - *age_category:* → the 'create a new attribute called' option, and set the *Label:* to `Age`
    - *ethnicity:* → the 'create a new attribute called' option, and set the *Label:* to `Ethnicity`
    - *grew_up:* → the 'create a new attribute called' option, and set the *Label:* to `Country`
    - *grew_up_region:* → the 'create a new attribute called' option, and set the *Label:* to `Region`
    - *grew_up_town:* → the 'create a new attribute called' option, and set the *Label:* to `Town`

- *languages_spoken:* → the 'create a new attribute called' option, and set the *Label:* to `Languages`

37. Click *import*.
    You should see a page with information about the import, including the columns that were ignored or added, and the number of participants that were added.

    To check the participant attributes really are now set:

38. Click the *participants* option on the menu.
    You will see a list of ten speakers, and page links at the bottom.

39. Pick a speaker (e.g. QB702_AnnaSoboleva) and click their *Attributes* button.
    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.

40. Click the *participant attributes* link on the menu.
    This will display a list of the participant meta-data fields.

41. Ensure that *searchable:* is set to 'Searchable' for the following attributes:
    - "gender"
    - "age_category"
    - "languages_spoken"

42. Click the *Save Changes* button in the top right corner.

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

Exercise 3 Basic 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.

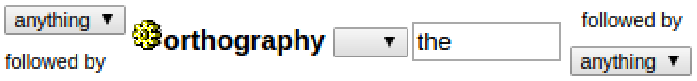
1. In LaBB-CAT, click on the *search* link on the menu.
   This takes you to a page entitled "filter", where you can list participants and filter them by their attributes. You can see various participant attributes listed across the page.

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 Micorsoft Excel and you find it doesn't open all the columns correctly:

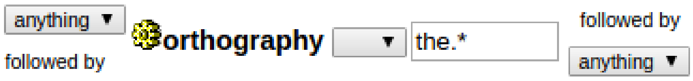
    (a) Create a new workbook in Excel.
    (b) Click the 'Data' tab.
    (c) On the "Get External Data" ribbon click 'From Text'.
    (d) Select the CSV file you downloaded.
    (e) Select 'Delimited' and click *Next*.
    (f) Ensure 'Comma' is the only delimiter ticked and click *Next*.
    (g) Click *Finish* 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. There are a few more uses for the CSV results files, which are dealt with in a separate tutorial...

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..."

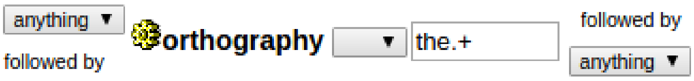
18. Change the *orthography* search text to `the.*` (i.e. after the word "the", append a full-stop and an asterisk.



The 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"



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.

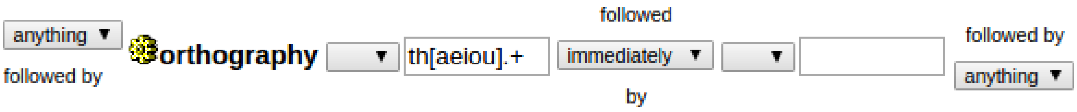
22. Now change your search by replacing the `e` in "the" with `[aeiou]` – so your search pattern will be `th[aeiou].+`
The square-brackets mean "any one of the things inside the brackets", so `[aeiou]` means "any vowel".

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 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*.



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.

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

Because we're searching by word orthography, you will have noticed that your searches for words starting with a vowel return words where the *spelling* starts with a vowel, but the *pronunciation* doesn't, e.g. "one", "once", etc. In order to search by pronunciation, we need to add a layer of pronunciation annotations. We'll do that in the next exercise...

Exercise 4 The CMU Dictionary and Cross Layer Searching

LaBB-CAT can be integrated with the CMU Pronouncing Dictionary, which is a free pronounciation dictionary of English maintained by the Speech Group in the School of Computer Science at Carnegie Mellon University. The pronunciations are based on American English, so are suitable for American English recordings.

It can also serve as a free alternative to the CELEX lexicon (which is based on British English), for those that have not purchased CELEX, although is less ideal for 'non-rhotic' varieties of English.

In this exercise you will:

- install the CMU Pronouncing Dictionary layer manager,

- use it to create new annotations for word pronunciations, and
- incorporate the new layers in more sophisticated searches.

The first thing we're going to to is install the CMU Pronouncing Dictionary layer manager...

1. Click the *layer managers* menu option.
   You will see a list of pre-installed layer managers, which are modules that perform automatic annotation tasks. The CMU Pronouncing Dictionary layer manager isn't pre-installed, because it is language-specific.
2. Click the *List of layer managers that are not yet installed* link near the bottom.
3. Find "CMU Pronouncing Dictionary" in the list, and click its *Install* button. You will see a progress bar while the layer manager loads the data from the dictionary file into the LaBB-CAT database. This will take a minute or so.
4. Once it's finished, you will see a new window open with information about the CMU Pronouncing Dictionary layer manager.
   Reading this information page, you will see some instructions on how to create a pronunciation annotation layer.

Now that we've installed the layer manager, we'll create a layer that contains word pronunciations. Follow the instructions on the information page to create a layer for word pronuncation, i.e.:

5. Click on the *word layers* option on the menu.
   You will see a list of existing word layers, including the *orthography* layer, the *lexical* layer, etc.
6. Scroll to the bottom and fill in the last, blank row:
   - *description:* `Pronunciation (CMUdict)`
   - *short description:* `phonemes`
   - *layer type:* 'Phonological'
   - *layer manager:* 'CMU Pronouncing Dictionary'
   - *alignment:* 'Not aligned'
7. Click *New* to add the layer.
   You will see the layer configuration form.
8. Set the *Encoding:* field to 'DISC', and the default values for everything else.
   If you're curious about what the configuration options do, the online help page when you are configuring the layer.
9. Click *Save*
   You will see a message asking you if you want generate the layer data now.
10. Click *Regenerate.*
    You will see a progress bar moving across the page while the annotations are being generated. When it is finished, you will see a message saying "Layer complete..."
11. Once the layer has finished generating, click the transcripts menu option, and open the first transcript in the list.
12. Tick your new *phonemes* layer.
    You will see that each word is tagged with a phonemic transcription. You will notice that the annotations are displayed using IPA symbols. However, the layer manager doesn't use IPA symbols directly, it actually uses the 'DISC' encoding for phonemes, which uses ordinary 'typewriter' characters (ASCII), and uses exactly one character per phoneme.
    The IPA symbols are being displayed by LaBB-CAT to provide a linguist-friendly representation of the phonemic transcription. But you can see the underlying DISC characters by selecting the 'ASCII' option on the layer in the transcript.
13. Select 'ASCII' on the phonemes layer, to see what the layer manager is actually producing.
    You may find that this is somewhat harder to read. It's similar to the 'SAMPA' system for encoding phonemes, but diphthongs are generally represented by digits, and various other characters are used to represent affricates, etc.

IPA	DISC	ARPABET		IPA	DISC	ARPABET	
p	p	P	pat	ɪ	I	IH	KIT
b	b	B	bad	ɛ	E	EH	DRESS
t	t	T	tack	æ	{	AE	TRAP
d	d	D	dad	ʌ	V	AH	STRUT
k	k	K	cad	ɒ	Q	AH	LOT
ɡ	g	G	game	ʊ	U	UH	FOOT
ŋ	N	NG	bang	ə	@		another
m	m	M	mad	iː	i	IY	FLEECE
n	n	N	nat	ɑː	#	AA	father
l	l	L	lad	ɔː	$	AO	THOUGHT
r	r	R	rat	uː	u	UW	GOOSE
f	f	F	fat	ɜː	3	ER	NURSE
v	v	V	vat	eɪ	1	EY	FACE
θ	T	TH	thin	aɪ	2	AY	PRICE
ð	D	DH	then	ɔɪ	4	OY	CHOICE
s	s	S	sap	əʊ	5	OW	GOAT
z	z	Z	zap	aʊ	6	AW	MOUTH
ʃ	S	SH	sheep	ɪə	7		NEAR
ʒ	Z	ZH	measure	ɛə	8		SQUARE
j	j	Y	yank	ʊə	9		CURE
x	x		loch	æ̃	c		timbre
h	h	HH	had	ɑ̃ː	q		détente
w	w	W	wet	æ̃ː	0		lingerie
tʃ	J	CH	cheap	ɒ̃ː	~		bouillon
dʒ	_	JH	jeep				
ŋ̩	C		bacon				
ɱ̩	F		idealism				
n̩	H		burden				
l̩	P		dangle				

Table 1: IPA to DISC Correspondences

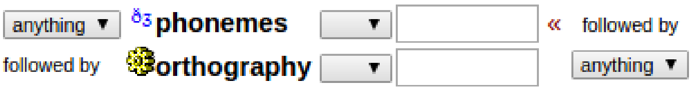
14. Select 'IPA' on the phonemes layer, to return to the IPA view of the layer.

It's nice to display the IPA symbols, but it's important to understand the DISC symbols (shown in Table 1), because they are what we have to use when searching on the phonemes layer, which we are going to try now.

As you may have seen on the layer configuration page, there is another possible representation of the pronunciations, called ARPABET; this is what is used in the original dictionary file published by CMU, and uses up to three uppercase characters per phoneme. While we're not using ARPABET in this exercise, you can use it if you like, and the ARPABET symbols are included in Table 1. In the table, you will see that there are gaps where no ARPABET version of the phoneme is shown; this means that the CMU Pronouncing Dictionary contains no entries that include that phoneme.

15. Go to the *search* page and select all speakers.

16. If it's not already ticked, tick the new *phonemes* layer and click *Set Search Matrix*.
Now you will see that our search matrix is two layers high by one word wide.



17. Search your new *phonemes* layer for words that start with `h` by entering the approriate regular expression in the *phonemes* box.
You will see that the results contain words that you might not expect, like "where", "which" and "when".

18. Click one of these unexpected results, to open the transcript.
You will see that, in the transcript, the pronunciation appears to start with /w/, not with /h/.

19. Click on the word and select the 'Edit' option on the menu that appears.
Now look for the *phonemes* layer. You will see that, in addition to the pronunciation that starts with /w/, there's another annotation that starts with /h/, which is invisible on the transcript.

These are all the possible phonemic transcriptions for the word, ordered most-frequent first. Only the first one is displayed in the transcript, but when you do searches, all of them are searched. This can result in unexpected matches like this, but it can be useful, as it ensures that when you search for a particular phonemic pattern, all possible tokens are returned, not just those that match on the most 'normal' transcription.

Now that we have phonemic transcripts, we can do a better job of the search we tried in the earlier exercise – "the" followed by a word starting with a vowel...

20. Go to the *search* page and select all speakers.
21. Create a search matrix that's two words wide, and includes the *orthography* and *phonemes* layers.
22. Type `the` in the first *orthography* box.
23. Click the second box on the *phonemes* layer, but don't enter anything in the box yet.
24. The box has a little symbol « to the right of it.
    Hover the mouse over it to see what it says, and then click it.
    You will see that a section opens with a bunch of phoneme symbols on it; clicking on a phoneme adds its DISC representation to the search box.
25. You could use the square-brackets `[` at the start of your pattern, and click all vowel symbols to add all possible vowels – Note that the vowels in the DISC representation extend beyond a, e, i, o, and u – you should add in all the vowels you see in the list that appears when you expand the IPA helper, including all the diphthongs.
    Alternatively, you can simply click the *VOWEL* link in the 'IPA helper', which will add all the DISC vowels for you, already enclosed in square-brackets.
26. Run the search and check that it's giving you what you expect. Notice that now there are no 'false positives' like "the one" that we were getting when searching by orthography alone.

Now that you've generated an annotation layer, and have seen how the search matrix works, you might want to try out some of the following searches, or invent some others:

- Words which have the DRESS vowel as the second phoneme
- Words ending with a front vowel, followed by words beginning with /p/ or /b/
- Words that begin with "k" in their spelling, but begin with the phoneme /n/
- Words that begin with "k" in their spelling, but *do not* begin with the phoneme /n/