# 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:

1. install the CMU Pronouncing Dictionary layer manager,
2. use it to create new annotations for word pronunciations, and
3. incorporate the new layers in more sophisticated searches.

## Install the CMU Dictionary

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

1. Select the *layer managers* menu option.
   You will see a list of pre-installed layer managers, which are modules that can perform automatic annotation tasks. The CMU Pronouncing Dictionary layer manager isn't pre-installed, because it is language-specific.
2. Follow the *List of layer managers that are not yet installed* link near the bottom.
3. Find *CMU Pronouncing Dictionary* in the list, and press its *Install* button.
4. Press *Install* on the resulting information page.
   This displays some further information about the layer manager, allowing you to upload an alternative version of the dictionary file.
   We be using the standard file that is included with the layer manager.
5. Press *Configure*.
   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.
6. Once it's finished, you will see a new window open with information about the CMU Pronouncing Dictionary layer manager.

## Annotate Words with Pronunciations

Now that we've installed the layer manager, we'll create a layer that contains word pronunciations.

7. Select the *word layers* option on the menu.
   You will see a list of existing word layers, including the *orthography* layer, the *lexical* layer, etc.

8. The column headings are also a form for defining a new word layer. Fill in the following details in this form:
    - **Layer ID:** `phonemes`
    - **Type:** *Phonological*
    - **Alignment:** *None*
    - **Manager:** *CMU Pronouncing Dictionary*
    - **Description:** `All possible phonemic transcriptions for each word.`
9. Press *New* to add the layer.
   You will see the layer configuration form.
10. Set the **Encoding** field to *CELEX DISC*, and the default values for everything else.

> **Tip**
>
> If you're curious about what the configuration options do, hover your mouse over each one to see further information about what the setting does.

11. Press *Set Parameters*.
    You will see a message asking you if you want (re)generate the layer data now.
12. Press *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*.
13. Once the layer has finished generating, click the transcripts menu option, and open the first transcript in the list.
14. 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.
15. 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.
16. Select *IPA* on the phonemes layer, to return to the IPA view of the layer.

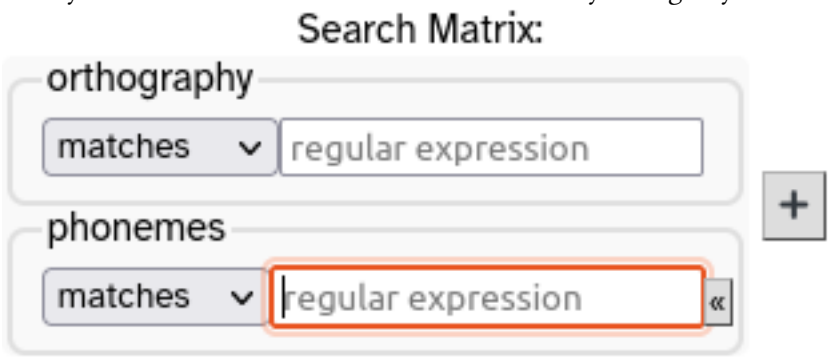
## Search Across Layers

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

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 the table. In the table, 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.

| IPA | DISC | ARPABET | | IPA | DISC | ARPABET | |
|---|---|---|---|---|---|---|---|
| p | p | P | **p**at | ɪ | I | IH | K**I**T |
| b | b | B | **b**ad | ɛ | E | EH | DR**E**SS |
| t | t | T | **t**ack | æ | { | AE | TR**A**P |
| d | d | D | **d**ad | ʌ | V | AH | STR**U**T |
| k | k | K | **c**ad | ɒ | Q | AH | L**O**T |
| g | g | G | **g**ame | ʊ | U | UH | F**OO**T |
| ŋ | N | NG | ba**ng** | ə | @ | [vowel ending in 0] | **a**nother |
| m | m | M | **m**at | iː | i | IY | FL**EE**CE |
| n | n | N | **n**at | ɑː | # | AA | f**a**ther |
| l | l | L | **l**ad | ɔː | $ | AO | TH**OU**GHT |
| r | r | R | **r**at | uː | u | UW | G**OO**SE |
| f | f | F | **f**at | ɜː | 3 | ER | N**UR**SE |
| v | v | V | **v**at | eɪ | 1 | EY | F**A**CE |
| θ | T | TH | **th**in | ɑɪ | 2 | AY | PR**I**CE |
| ð | D | DH | **th**en | ɔɪ | 4 | OY | CH**OI**CE |
| s | s | S | **s**ap | əʊ | 5 | OW | G**OA**T |
| z | z | Z | **z**ap | ɑʊ | 6 | AW | M**OU**TH |
| ʃ | S | SH | **sh**eep | ɪə | 7 | | N**EAR** |
| ʒ | Z | ZH | mea**s**ure | ɛə | 8 | | SQU**ARE** |
| j | j | Y | **y**ank | ʊə | 9 | | C**URE** |
| x | x | | lo**ch** | æ | c | | t**i**mbre |
| h | h | HH | **h**ad | ɑ̃ː | q | | dét**en**te |
| w | w | W | **w**et | æ̃ː | 0 | | l**in**gerie |
| ʧ | J | CH | **ch**eap | ɒ̃ː | ~ | | bouill**on** |
| ʤ | _ | JH | **j**eep | | | | |
| ŋ̩ | C | | bac**on** | | | | |
| m̩ | F | | idealis**m** | | | | |
| n̩ | H | | burd**en** | | | | |
| l̩ | P | | dang**le** | | | | |

17. Select the *search* option on the menu, which allows you to search all participants by default.
18. If it's not already ticked, tick the new *phonemes* layer.
    Now you will see that our search matrix is two layers high by one word wide.




19. Search your new *phonemes* layer for words that start with `h` by entering the appropriate regular expression in the *phonemes* box.
    You will see that the results contain words that you might not expect, like "where", "which" and "when".
20. 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/.
21. Click on the word and select the bottom *Edit* option on the menu that appears.
    This opens a small window that displays all annotations on that word token.
    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…

22. Go to the *search* page.
23. Create a search matrix that's two words wide, and includes the *orthography* and *phonemes* layers.
24. Type `the` in the first *orthography* box.
25. Click the second box on the *phonemes* layer, but don't enter anything in the box yet.
    The box has a little symbol « to the right of it.
26. 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.
27. 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 'phoneme symbol selector', which will add all the DISC vowels for you, already enclosed in square-brackets.

28. Be sure to append the 'any vowel' regular expression with `.*` to ensure the search matches words that have phonemes after the initial vowel
29. 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/