# 7b. CMU Pronouncing Dictionary

LaBB-CAT can be integrated with the CMU Pronouncing Dictionary, which is a free pronunciation 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
- Incorporate the new layers in more sophisticated searches

---

The first thing we're going to do is install the CMU Dict layer manager…

(1) Select the *layer managers* menu option.
(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, *Install* again, and then the *Configure* 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 page with information about the CMU Pronouncing Dictionary layer manager.

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

(5) Add a *word layer* managed by the *CMU Pronouncing Dictionary* for word pronunciation - i.e.:

    - **Layer ID:** `phonemes`
    - **Type:** *Phonological*
    - **Alignment:** *None*
    - **Manager:** *CMU Pronouncing Dictionary*
    - **Description:** `CMU Pronouncing Dictionary pronunciations`
    - …configured with the **Encoding:** field set 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 option to see a 'tool tip' that describes what the option is for.

(6) Once the layer has finished generating, select the *transcripts* menu option, and find and open *NB926_IsobelleDoig.eaf*.
(7) 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.

(8) 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. Diphthongs are generally represented by digits, and various other characters are used to represent affricates, etc.

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.

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 the table. 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.

| 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 o] | **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 | | | | |
| n̩ | C | | bac**on** | | | | |
| m̩ | F | | idealis**m** | | | | |
| n̩ | H | | burd**en** | | | | |
| l̩ | P | | dang**le** | | | | |

In the transcript, you may notice there are gaps in the layer - i.e. words that are not tagged with a pronunciation.

For example, around the middle of the transcript, the word “compactums” is not tagged, because the CMU Pronouncing Dictionary has no entry for that word.

There are various possible solutions for this, but one is to tag word tokens with their pronunciations directly in the transcript. This has been done in the case of “compactums”; manual pronunciation tags are saved on the *pronounce* layer

(9) Scroll to the top of the transcript, un-tick the *phonemes* layer and tick the *pronounce* layer.
(10) When the transcript re-loads to show the *pronounce* layer tags, find "compactums" again.

You will see it has been tagged with an annotation labelled "kəmpæktəmz", which was manually added by the transcriber of the transcript, in the original ELAN file.

We want all pronunciations to be present on the *phonemes* layer, which is currently managed by the CMU Pronouncing Dictionary layer manager. LaBB-CAT allows layers to have more than one layer manager, however; a layer can have a main layer manager, and a number of 'auxiliary' managers that perform extra annotation tasks.

We are going to add an auxiliary layer manager to the *phonemes* layer, which will copy any *pronounce* annotations it finds to the *phonemes* layer. This will fill in the gaps in the CMU Pronouncing dictionary, at least for the tokens that have manual *pronounce* tags.

(11) Select the *word layers* option on the menu.
(12) On the *phonemes* layer row, there are a number of buttons on the right, including one with a icon. Hover your mouse over this button to see what it does, and then press it.
You will see a page explaining that will copy any manually tagged pronunciations from the from the *pronounce* layer into the *phonemes*.
(13) Pres *yes* to continue.
You will see a progress bar while the auxiliary layer manager copies the *pronounce* annotations to the *phonemes* layer.
(14) When it's finished, select the *transcripts* menu option, and open *NB926_IsobelleDoig.eaf* again.
(15) Tick the *phonemes* layer.
(16) Find the word "compactums" in the transcript.

You will see it now has a *phonemes* tag, just like the rest of the word tokens.

(17) Select the *search* option from the menu.
(18) Search your new *phonemes* layer for words that start with *h*

You will see that the results contain words that you might not expect, like "where", "which" and "when".

(19) 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/.

(20) 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. 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 we're going to try to do a search for the word "the" followed by a word that starts with schwa.

(21) Select the *search* option from the menu.
(22) Create a search matrix that's two words wide, and includes the *orthography* and *phonemes* layers.
(23) Type *the* in the first *orthography* box.
(24) Click the second box on the *phonemes* layer, but don't enter anything in the box yet.
(25) The box has a button 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.
(26) Find the schwa symbol *ə* and click it.
You will see that an `@` symbol appears in the box.
`@` is the DISC symbol for /*ə*/, so in order to search for schwa, we have to use it in our search pattern.
(27) We want words that start with schwa, so type `.*` after the `@` symbol.
(28) Click *Search*.

You will see that some of the words being matched are words that you might not normally think start with a schwa. LaBB-CAT is matching words against *all their possible phonemic transcriptions*, so if the CMU dictionary has multiple possible pronunciations for a word, and one of them starts with schwa, it will be matched.

You can check this by clicking on a match, and then clicking on the word in the transcript and selecting *Edit*, which displays all the annotations for the given token.

If you check the table above, you will see that *ə* has no specific representation in ARPABET. This means that no CMU Pronouncing Dictionary pronunciations include schwa explicitly. Instead, 'unstressed' versions of other vowels are used. For example, the word "transcription" is transcribed `T R AE2 N S K R IH1 P SH AH0 N` in the original dictionary file; the final vowel *AH* is the 'STRUT' vowel, and the *0* means it's 'unstressed'. The layer manager translates this to DISC as `tr{nskrIpS@n`.

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

(29) Change your search so that, instead of just `@` at the beginning of the word, it matches any vowel.

> **Tip**
>
> You could use the square-brackets `[]` at the start of your pattern, and type all vowel symbols inside them – 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.

(30) 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 a few different layers, 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
- The word "the" followed by a word beginning with the phoneme /k/
- 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/