

Exercise 6 Integrating with the CMU Pronouncing Dictionary

LaBB-CAT can be integrated with the CMU Pronouncing Dictionary, which is a free pronouncing 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. Click the *layer managers* menu option.
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.

5. Follow the instructions on the information page to create a layer for word pronunciation – i.e.:
 - *description*: Pronunciation (CMUdict)
 - *short description*: phonemes
 - *layer type*: ‘Phonological’
 - *layer manager*: ‘CMU Pronouncing Dictionary’
 - *alignment*: ‘Not aligned’
 - ... configured with the *Encoding*: field set 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.
6. Once the layer has finished generating, click the transcripts menu option, and open the first transcript in the list.
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.

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
g	g	G	game	ʊ	U	UH	FOOT
ŋ	N	NG	bang	ə	@		another
m	m	M	mad	i:	i	IY	FLEECE
n	n	N	nat	ɑ:	#	AA	father
l	l	L	lad	ɔ:	\$	AO	THOUGHT
r	r	R	rat	u:	u	UW	GOOSE
f	f	F	fat	ɜ:	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	ã	c		tĩmbre
h	h	HH	had	ã:	q		détente
w	w	W	wet	ã:	0		lĩngerie
tʃ	J	CH	cheap	õ:	~		bouillon
dʒ	—	JH	jeep				
ŋ	C		bacon				
ɪ	F		idealism				
ɪ	H		burden				
ɪ	P		dangle				

Table 1: IPA to DISC Correspondences

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.

9. Go to the *search* page and select all speakers.
10. 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”.
11. 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/.
12. 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 we're going to try to do a search for the word “the” followed by a word that starts with schwa.

13. Go to the *search* page and select all speakers.
14. Create a search matrix that's two words wide, and includes the *orthography* and *phonemes* layers.
15. Type **the** in the first *orthography* box.
16. Click the second box on the *phonemes* layer, but don't enter anything in the box yet.
17. 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.
18. Find the schwa symbol ə and click it.
You will see that a @ 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.
19. We want words that start with schwa, so type .* after the symbol.
20. Click *Search*.
You will see that, surprisingly, no results are returned. Why is this?

If you check Table 1, you will see that ə has no representation in ARPABET. This means that no CMU Pronouncing Dictionary pronunciations include schwa. Instead, these include ‘unstressed’ versions of other vowels. 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{nskrIpsVn, which includes V as the final vowel instead of ə as you might expect.

However, 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...

21. Change your search so that, instead of just **the** at the beginning of the word, it matches any vowel.
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 ‘IPA helper’, which will add all the DISC vowels for you, already enclosed in square-brackets.

22. 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/