Phonemic Tagging with 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.

Integrating this lexicon with LaBB-CAT is achieved with the "CMU Dictionary Layer Manager". As CMU has kindly granted permission to freely distribute the dictionary file, you don't need to download the file from the CMU site; it's included in the layer manager that you will install.

Install the Layer Manager

First, the *CMU Pronouncing Dictionary* layer manager module must be installed:

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, then *Install* again.
4. Press *Configure* to install the default version of the dictionary.

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.

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.

Create an Annotation Layer

When tagging words with their pronunciations, the *CMU Pronouncing Dictionary* layer manager allows two possible options for the symbols used in the transcriptions:

CMU ARPAbet e.g. `T R AE2 N S K R IH1 P SH AH0 N` – this is the original encoding used in the CMUdict file, and it's also the symbol set used for the Penn Aligner pre-trained models. Each phoneme may be multiple characters long, and the transcriptions use a space delimiter between each phoneme, which can make searching using regular expressions more complicated.

CELEX DISC e.g. `tr{nskrIpS@n` – this is one of the encodings used by the CELEX lexicon. Each phoneme is represented with exactly one ASCII character, so pattern matching with regular expressions is somewhat easier.

Before creating the layer to tag each word with its pronunciation(s), you must decide which encoding the layer will use.

To create a new layer with CMUdict annotations:

1. Select on the *word layers* menu option - this will display a list of all the word layers you already have in the database.
2. At the top of the list, there's a blank forARPAm for creating a new layer - fill this form in:
    - **Layer ID**: `phonemes`
    - **Type**: *Text* if you're using the *CMU ARPAbet* encoding, or *Phonological* if you're using the *CELEX DISC* encoding.
    - **Manager**: *CMU Pronouncing Dictionary*
    - **Alignment**: *None* (as these are simply tags on the orthographic words)
    - **Generate**: *Always*
    - **Description**: *All possible phonemic transcriptions, according to CMUdict*
3. Press the *New* button to create the layer
   You will see a form that allows you to specify various options, including the *Encoding* to use for transcriptions.
4. Ensure the *Encoding* is set to the option you prefer.
5. Press *Set Parameters*
6. Press *Regenerate*.

LaBB-CAT will then generate annotations for all the transcripts you already have in your database. If you have a lot of data, this may take a while.

From now on, when you upload a new transcript, the CMUdict annotations will automatically be generated for it.