Phonemic Tagging using a Plain Text File

The *Flat Lexicon* layer manager annotates words with data from a dictionary loaded from a plain text file (e.g. a CSV file).

The dictionary file you supply may contain multiple fields, and multiple entries per word. It might include:

- word orthography
- lemma
- part-of-speech
- pronunciation
- frequency

…or any other "type" data you like.

Getting a dictionary file

What dictionary file you want depends on what you want to annotate. For pronunciations, you might download some standard dictionary for your target language, such as Unisyn, the CMU Pronouncing dictionary, CELEX, etc. (although there are also specialised layer managers for these particular lexicons). Frequency lists include CELEX, SubtlexUS, and Adam Kilgarriff's BNC Frequency Lists.

Alternatively, you might have, or prepare, your own dictionary containing pronunciations, lemmata, etc. All you need is a CSV file with a column that includes the word orthography, and other columns that include the pronunciation and any other information you may have.

	A	B	C	D
1	words	X-SAMPA	DISC	ARPABET
2	ā	a:	a	AA1
3	ā'a	a:?'a	a?a	AA0 AH1
4	āfei	a:f'ei	af1	AA0 F EH1 IH0
5	āata	a:'ata	aata	AA0 AH1 T AH0
6	kī	k'i:	ki	K IY1
7	kīkīvoi	ki:k'i:voi	kikiv3	K IY0 K IY1 V AO0 IH0
8	kītaki	ki:t'aki	kitaki	K IY0 T AH1 K IH0
9	toko	t'oko	toko	T EH1 K EH0

Figure 1: Example of a custom pronunciation dictionary

NB LaBB-CAT assumes that the text file uses ASCII or UTF-8 character encoding. If your dictionary file uses another encoding (e.g. "Western" or ISO-8859, you will need to re-save the file using UTF-8 (in many text editors, the character encoding is an option available when you select "Save As…" from the "File" menu).

Installing a dictionary file

Once you have a CSV or other text file, you need to upload it into LaBB-CAT:

1. Select the *layer managers* option in the LaBB-CAT menu.
2. Find "Flat Lexicon Tagger" in the list and press its *Extensions* button.
3. Press *Choose File* and select your dictionary file.
4. You many decide to change the default "Name" that the lexicon will have.
5. The default file structure options will probably be correct, but you may change them if you need to - see the page's online help for details.
6. Press *Load*.

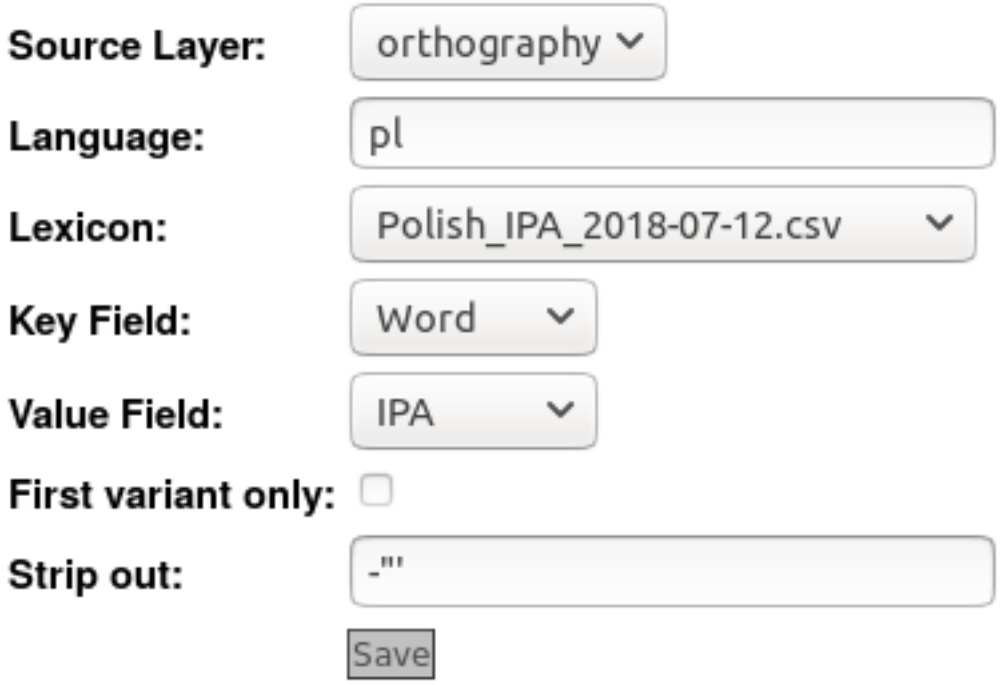
You can upload as many dictionaries as you like. Once you have at least one dictionary, you can configure a word layer to lookup the resulting lexicons, by selecting "Flat Lexicon Tagger" as the layer's layer manager.

Creating a Phonemes Layer

To create a new layer with annotations from your dictionary:

1. Select the *word layers* option on the menu - 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 form for creating a new layer - fill this form in:
    - **Layer ID**: enter a one- or two-word description - e.g. phonemes
    - **Type**: If your dictionary uses CELEX DISC symbols that are not space delimited, select Phonological, otherwise (e.g. space-delimited IPA or ARPABET pronunciations) select Text
    - **Alignment**: select None (as these are simply tags on the orthographic words)
    - **Manager**: select Flat Lexicon Tagger
    - **Generate**: select Always
    - **Description**: enter a description of the layer - e.g. Pronunciation (text-file)
3. Press the *New* button to create the layer.
   You will see the layer configuration page. Check the online help for explanations of all options, but at least:
4. Ensure the *Source Layer* is *orthography*

5. Select the desired Lexicon from the list (these relate to the file or files you uploaded above).



6. Press *Save*
7. Press *Regenerate*.
   You will see a progress bar while the layer manager annotates all the transcripts that have already been uploaded.

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.

Once this is finished, be sure to open a transcript and tick the new phonemic tagging layer you just added, and make sure that each word is tagged with a corresponding pronunciation.

From now on, when you upload a new transcript, annotations will automatically be generated by lookup up your lexicon.