Phonemic Tagging: Multilingual Corpora

If the speech corpus includes data in more than one language, it is possible to ensure that the utterances are phonemically tagged in a way that's sensitive to the language of the specific utterance.

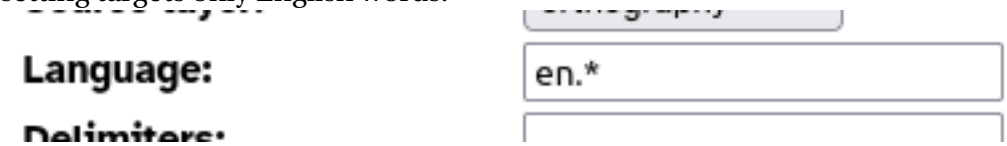
The layer manager modules that phonemically transcribe the data can be configured to annotate only words that are in the language targeted for that specific module, using the language code (e.g. "mi" for Te Reo Māori, "en" for English, "en-NZ" for New Zealand English, etc.).

Each annotation layer is usually managed by a single layer manager, but it's possible to have extra 'Auxiliary Layer Managers' configured for each layer. So you can have a single *phonemes* layer that contains all phonemic transcriptions, regardless of the language of the data; e.g. you might have the layer with

- the CELEX English layer manager as the primary layer manager, targeting only English utterances, plus
- the CELEX German layer manager as an auxiliary, targeting only German utterances, and
- the Character Mapper layer manager as an auxiliary, configured to target only utterances in Te Reo Māori with orthography-to-phonology mappings.

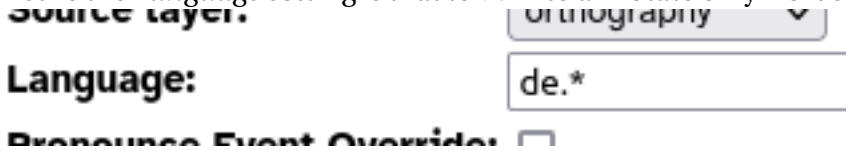
In order to set this up:

1. In LaBB-CAT, select the *word layers* option on the menu.
2. Add a *phonemes* layer that's managed by the CELEX English layer manager.
3. Set the layer configuration as required, ensuring that in its configuration, the *Language* setting targets only English words:



4. Select the *word layers* option on the menu again.
5. On the *phonemes* layer, press the 'Other configurations' icon:
6. Fill in the *description* box as `German pronunciation`
7. For the *layer manager* select the *CELEX German* option.
8. Press the *New* button.
9. Press *Configure*
10. Configure the layer as required.

11. Ensure the *Language* setting is that to `de.*` to annotate only words that are in German:



12. Select the *word layers* option on the menu again.
13. On the *phonemes* layer, press the 'Other configurations' icon:
14. Fill in the blank *description* box as `Te Reo Māori pronunciation`
15. For its *layer manager* select the *Character Mapper* option.
16. Press the *New* button.
17. Press *Configure*
18. Configure the layer as required.
19. Ensure the *Language* setting is that to `mi` to annotate only words that are in Māori:

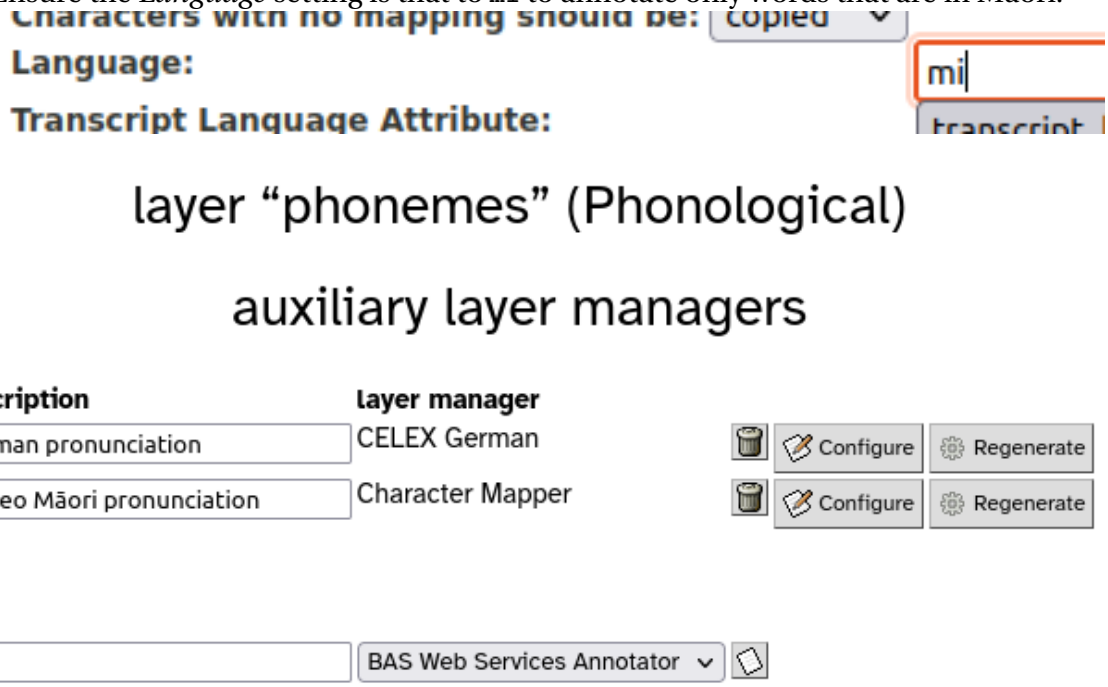


Figure 1: The 'phonemes' layer's two auxiliary configurations

When the layer is generated, first the main configuration will generate annotations (i.e. CELEX English phonology), and then the auxiliary configurations will be run, in alphabetical order. As long as each has a different language targeted, they will each annotate different word tokens.

LaBB-CAT has three mechanisms for determining the language of each word token in the

corpus:

1. If the word is enclosed in an annotation on the *language* phrase layer, then the annotation's label determines the language of that token. The language phrase layer is a time-span layer that allows spans of words to be marked as being in a specific language.

> **Tip**
>
> How these manual annotations are added depends on the transcription tool; e.g.
>
> - *Transcriber* has a mechanism specifically for this, and for ELAN transcripts, and
> - LaBB-CAT supports a language-tagging transcription convention which can achieve the same thing.

2. Otherwise, the transcript's *language* transcript attribute is used to determine the language.
3. If the *language* transcript attribute is unset, then the *language* of the corpus the transcript is in is used to determine the language.

Using these mechanisms, it's possible to ensure that each token is labelled with the correct phonemic transcription, even if the corpus contains multiple languages, and even if there are multiple languages within the same transcript.

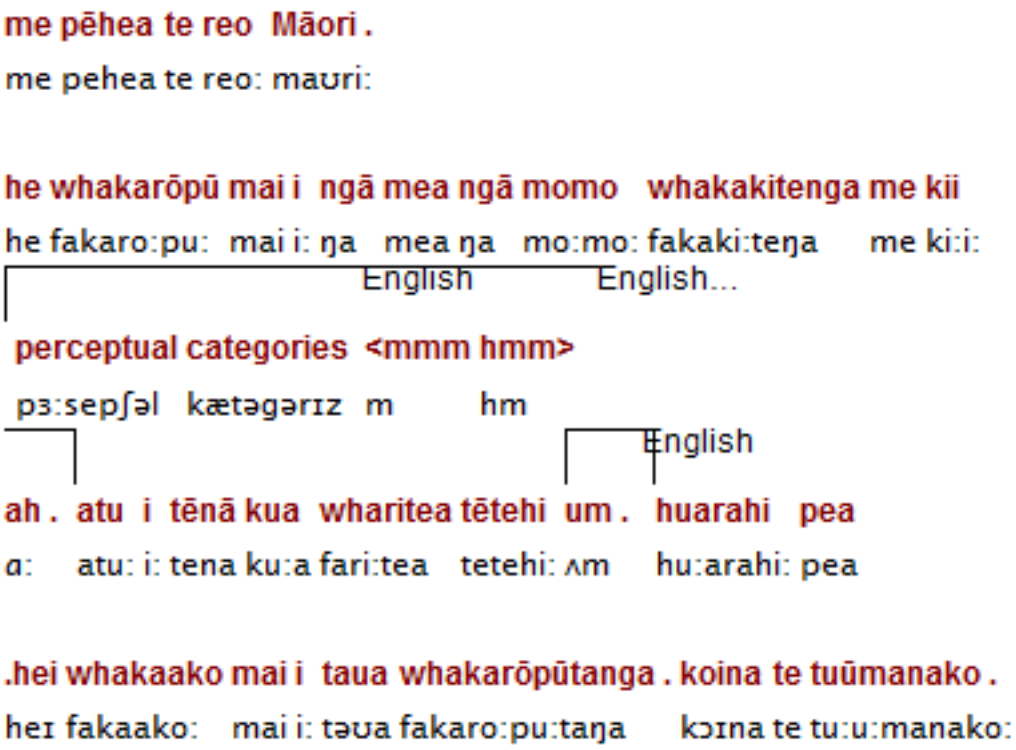


Figure 2: A transcript in Te Reo Māori, with English words annotated on the language phrase layer