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 Mapperg layer manager as an auxiliary, configured to target only utterances in Te Reo Māori with orthography-to-phonology mappings.

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 meta layer is a time-span layer that allows spans of words to be marked as being in a specific language.
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.

me pēhea te reo Māori .
me pehea te reo: mauri:

he whakarōpū mai i ngā mea ngā momo   whakakitenga me kii
he fakaro:pu: mai i: ŋa   mea ŋa   mo:mo: fakaki:teŋa   me ki:i:
English   English...

perceptual categories <mmm hmm>
pɜ:sepʃəl   kætəgərɪz m   hm
                                        English

ah . atu i tēnā kua wharitea tētehi um . huarahi   pea
ɑ: atu: i: tena ku:a fari:tea   tetehi: ʌm   hu:arahi: pea

.hei whakaako mai i taua whakarōpūtanga . koina te tuūmanako .
heɪ fakaako: mai i: təua fakaro:pu:taŋa   kɔɪna te tu:u:manako:

Figure 1: A transcript in Te Reo Māori, with English words annotated on the language phrase layer