Creating or modifying a dictionary using a g2p model

A dictionary is a necessary component for forced alignment, along with an appropriate acoustic model, and it has to contain all at least all the words in the transcriptions (you can add pronunciation variations, of course). This can be done using a large predefined dictionary (must be compatible with the acoustic model, of course), or by writing one by yourself. In this case, it has to follow the structure of the dictionary and the phone set.

Another option would be to use a g2p (grapheme-to-phoneme) model to create a dictionary. The MFA provides also this functionality and multiple pre-trained g2p models for different languages (see here).

This notebook shows how to use these models to create a prounciation dictionary. First, we have to download a pre-trained g2p model. We will use a g2p model for French for illustration.

```
In [1]: ! mfa model download g2p french_mfa
```

This g2p model is trained on the french_mfa pronunciation dictionary and will generate (possible) pronunciations based on this dictioinary.

We have to create now a simple text file, that contains one word per row. I entered here the french word articulerai (the missing word from the validation tutorial) as well as some other words (mostly non-words).

articulerai lonteaux oster mantau Hans

These words are saved in wordlist.txt . Pronunciations can be generated by entering the following command:

```
In [2]: # mfa g2p PATH_TO_YOUR_WORDLIST PATH_TO_G2P_MODEL PATH_TO_OUTPUT_DICTIO
```

In the present case, I entered the following command, using the french_mfa g2p model, therefore no path has to be specified:

```
In [3]: ! mfa g2p --clean ~/Workspace/corpora/wordlist.txt french_mfa ~/Workspa
```

```
INFO - Generating pronunciations...
100%|█████████████████████████████████████████████| 5/5 [00:01<00:00,
4.45it/s]
INFO - Done! Everything took 1.787 seconds
```

The resulting dictionary has the structure necessary for the MFA and has following entries:

articulerai a ʁ t i k y l ʁ e

lonteaux l ɔ̃ t o

oster ɔ s t e

oster ɔ s t ɛ ʁ

mantau m ɑ̃ t o

Hans a n

Hans ɑ̃

The pronunciation of the first word corresponds to the transcription we already entered in the validation tutorial. The other entries seem to follow the french phonology, but we have some alternations present. Especially regarding the name Hans we see that the generated transcriptions do not match the way this name is pronounced. The g2p dictionary creation works perfectly, but the reason for this weird transcription lies in the french_mfa dictionary that was used for training. Looking into the dicionary more closely, we see 1) that there is no entry where the ortographic word starts with <h> which is also present in the transcription, 2) there are not many items in which the ortographic word and the transcription end with <s> and [s] respectively; this makes a word-final [s] unlikely in these cases, 3) most orthographic <an> sequences correspond to either [ɑ̃] or [an] - this is why we got these transcriptions.

Although this is a highly constructed/special case, it is always good to check whether the transcriptions are correct/are what we need, and modify them if necessary.

The g2p dictionary generation can be a nice tool for creating a new dictionary or modifying an existing one. In the latter case, we can enter all the words that are in the transcriptions but not in the dictionary into a simple text file, generate a dictionary and copy & paste the content into an existing one.

g2p models may not be available for every language covered by the MFA. In some cases, you may have an MFA dictionary but no g2p model. In these cases, you can create a new g2p model (similar to the training of a new acoustic model) and use it with the command above. The command for training a new g2p model is:

In [4]:

```
# mfa train_g2p PATH_TO_YOUR_DICTIONARY PATH_TO_THE_G2P_MODEL_OUTPUT
```

Similar to the training of a new acoustic model, it is best to enter the name of your g2p model and a .zip suffix at the end of the path.