Validating a corpus

If we want to apply the MFA on own or external data using pretrained models and dictionaries, it is always good to check wether al the words in our transcription files are also included in the dictionary. If this is not the case, we have to modify the dictionary and add the missing words. This notebook shows how to use the validation function of the MFA.

The MFA provides a function that will check if sound files or transcriptions are missing and whether all words in our transcriptions have also an entry in a pronunciation dictionary. The command is:

In [1]:
```
# mfa validate --clean PATH_TO_YOUR_DATA PATH_TO_YOUR_DICTIONARY
```

In order to test this function, we will use the Corpus_validation_dataset.zip and the french_mfa_dict_modified.txt from the Google Drive folder. THe zip archive contains wav and lab files from the first part of the SIWIS French Speech Synthesis corpus. Create a corpus folder and unpack the SIWIS data into it. Place the dictionary file somewhere you want. In this example I created a folder french_data in a Workspace/corpora/ directory, and the dictionary is in Workspace/corpora/. The full specified command is:

In [2]:
```
! mfa validate --clean ~/Workspace/corpora/french_data/ ~/Workspace/cor
```

```
INFO - Setting up corpus information...
INFO - Loading corpus from source files...
100%|████████████████████████████████████████| 120/120 [00:01<00:00, 11
4.13it/s]
INFO - Found 1 speaker across 120 files, average number of utterances p
er
                speaker: 120.0
INFO - Initializing multiprocessing jobs...
WARNING - Number of jobs was specified as 3, but due to only having 1 s
peakers,
                  MFA will only use 1 jobs. Use the --single_speaker f
lag if
                  you would like to split utterances across jobs regar
dless of
                  their speaker.
INFO - Normalizing text...
100%|████████████████████████████████████████| 120/120 [00:07<00:00, 1
5.45it/s]
INFO - Creating corpus split for feature generation...
100%|████████████████████████████████████████| 240/240 [00:01<00:00, 22
8.46it/s]
INFO - Generating MFCCs...
100%|████████████████████████████████████████| 120/120 [00:02<00:00, 4
9.74it/s]
INFO - Calculating CMVN...
INFO - Generating final features...
100%|████████████████████████████████████████| 120/120 [00:01<00:00, 11
1.09it/s]
INFO - Creating corpus split with features...
100%|████████████████████████████████████████| 120/120 [00:01<00:00, 6
0.26it/s]

INFO - ******
INFO - Corpus
INFO - ******

INFO -   120 sound files
INFO -   120 text files
INFO -   1 speakers
INFO -   120 utterances
INFO -   393.680 seconds total duration

INFO -   Sound file read errors
INFO -   ======================

INFO -     There were no issues reading sound files.

INFO -   Feature generation
INFO -   ==================

INFO -     There were no utterances missing features.

INFO -   Files without transcriptions
INFO -   ============================

INFO -     There were no sound files missing transcriptions.

INFO -   Transcriptions without sound files
INFO -   ==================================
```

```
INFO -     There were no transcription files missing sound files.


INFO - **********
INFO - Dictionary
INFO - **********

INFO -   Out of vocabulary words
INFO -   ========================

INFO -     There were no missing words from the dictionary. If you plan
on using
                the a        model trained on this dataset to align oth
er
                datasets in the future, it is        recommended that t
here be
                at least some missing words.


INFO - ********
INFO - Training
INFO - ********

INFO - Initializing training for monophone...
INFO - Compiling training graphs...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
2.90it/s]
INFO - Generating initial alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 9
9.83it/s]
INFO - Initialization complete!
INFO - monophone - Iteration 1 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:03<00:00, 3
1.55it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
7.35it/s]
INFO - monophone - Iteration 2 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 6
9.65it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
7.60it/s]
INFO - monophone - Iteration 3 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
4.06it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
6.57it/s]
INFO - monophone - Iteration 4 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
8.55it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
6.40it/s]
INFO - monophone - Iteration 5 of 40
```

```
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
9.97it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
7.11it/s]
INFO - monophone - Iteration 6 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
9.88it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
7.45it/s]
INFO - monophone - Iteration 7 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
9.68it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
6.89it/s]
INFO - monophone - Iteration 8 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
9.56it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
6.90it/s]
INFO - monophone - Iteration 9 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
9.34it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
4.77it/s]
INFO - monophone - Iteration 10 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
9.22it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
6.74it/s]
INFO - monophone - Iteration 11 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
7.03it/s]
INFO - monophone - Iteration 12 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
8.39it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
6.69it/s]
INFO - monophone - Iteration 13 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
6.39it/s]
INFO - monophone - Iteration 14 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
7.30it/s]
```

```
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
6.16it/s]
INFO - monophone - Iteration 15 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
6.91it/s]
INFO - monophone - Iteration 16 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
8.88it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
6.16it/s]
INFO - monophone - Iteration 17 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
6.33it/s]
INFO - monophone - Iteration 18 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
5.14it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
5.54it/s]
INFO - monophone - Iteration 19 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
5.85it/s]
INFO - monophone - Iteration 20 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
7.99it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
3.61it/s]
INFO - monophone - Iteration 21 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
5.69it/s]
INFO - monophone - Iteration 22 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
5.84it/s]
INFO - monophone - Iteration 23 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
7.49it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
4.86it/s]
INFO - monophone - Iteration 24 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
5.72it/s]
INFO - monophone - Iteration 25 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
5.62it/s]
INFO - monophone - Iteration 26 of 40
```

```
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
7.21it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
4.91it/s]
INFO - monophone - Iteration 27 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
5.25it/s]
INFO - monophone - Iteration 28 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
5.05it/s]
INFO - monophone - Iteration 29 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
7.01it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
4.61it/s]
INFO - monophone - Iteration 30 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
4.63it/s]
INFO - monophone - Iteration 31 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
4.68it/s]
INFO - monophone - Iteration 32 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
6.31it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
4.00it/s]
INFO - monophone - Iteration 33 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
3.98it/s]
INFO - monophone - Iteration 34 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
4.06it/s]
INFO - monophone - Iteration 35 of 40
INFO - Generating alignments...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 8
6.79it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
4.40it/s]
INFO - monophone - Iteration 36 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
4.79it/s]
INFO - monophone - Iteration 37 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████| 120/120 [00:01<00:00, 10
4.80it/s]
INFO - monophone - Iteration 38 of 40
```

```
INFO - Generating alignments...
100%|████████████████████████████████████████| 120/120 [00:01<00:00, 8
6.82it/s]
INFO - Accumulating statistics...
100%|████████████████████████████████████████| 120/120 [00:01<00:00, 10
2.34it/s]
INFO - monophone - Iteration 39 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████████| 120/120 [00:01<00:00, 10
4.53it/s]
INFO - monophone - Iteration 40 of 40
INFO - Accumulating statistics...
100%|████████████████████████████████████████| 120/120 [00:01<00:00, 10
4.68it/s]
INFO - Training complete!
INFO - Compiling training graphs...
100%|████████████████████████████████████████| 120/120 [00:01<00:00, 10
3.16it/s]
INFO - Generating alignments...
100%|████████████████████████████████████████| 120/120 [00:01<00:00, 8
2.51it/s]
INFO - Accumulating transition stats...
100%|████████████████████████████████████████| 120/120 [00:01<00:00, 10
9.45it/s]
INFO - Finished accumulating transition stats!
  0%|                                                    | 0/120 [00:00
<?, ?it/s]INFO - Collecting phone and word alignments from monophone_al
i lattices...
100%|████████████████████████████████████████| 120/120 [00:03<00:00, 3
2.73it/s]
INFO - Beginning phone LM training...
INFO - Collecting training data...
100%|████████████████████████████████████████| 120/120 [00:01<00:00, 11
4.14it/s]
INFO - Training model...
INFO - Completed training in 115.4295105934143 seconds!
INFO - Done! Everything took 177.944 seconds
```

After you run this command, we will get a lot of information about our corpus at the beginning of the output under Corpus. This is general information about the corpus such as the number of utterances, total duration of all files, if sound files are borken and much more. Under Dictionary you will find information about out of vocabulary words, these are words in our transcriptioni that are missing in the dictionary. In case there are out of vocabulary words, you can open the utterance_oovs.txt, that was placed in the MFA directory and a newly created folder that has the same name as the folder of your corpus. This file provides information about the words that are missing and in which file we can find it.

When we open the utterance_oovs.txt, we will see the following entry:

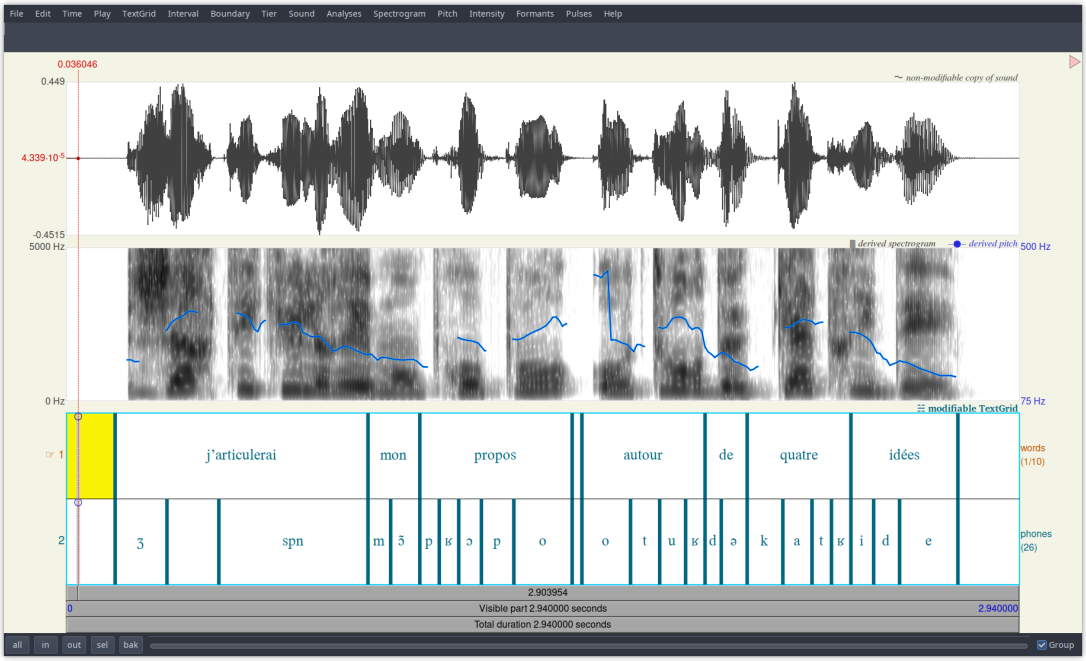
neut_parl_s01_0110, french_data: 0.0-2.94: a, r, t, i, c, u, l, e, r, a, i

In this case, the word articulerai is missing and it occures in the utterance neut_parl_s01_s0110. If we open the dictionary and search for this word, we will see that it is not included.

We can anyway try to align the data:

In [ ]:
```
! mfa align --clean ~/Workspace/corpora/french_data/ ~/Workspace/corpor
```

If we open the french_data_textgrids folder, we see that all of the 120 utterances were aligned, with the typical performance (based on the dictionary that does not include all possible pronunciation variations, of course). If we open the TextGrid with the missing word, we will see that there are no phone alignments for articulerai. Instead, we see spn which is a label for an unrecognized word.



We can solve this problem by entering an entry for articulerai manually into the dictionary:

articuler a ʁ t i k y l e /*articulerai a ʁ t i k y l e ʁ e*/ articulo a ʁ t i k y l o articulé a ʁ t i k y l e articulée a ʁ t i k y l e articulées a ʁ t i k y l e

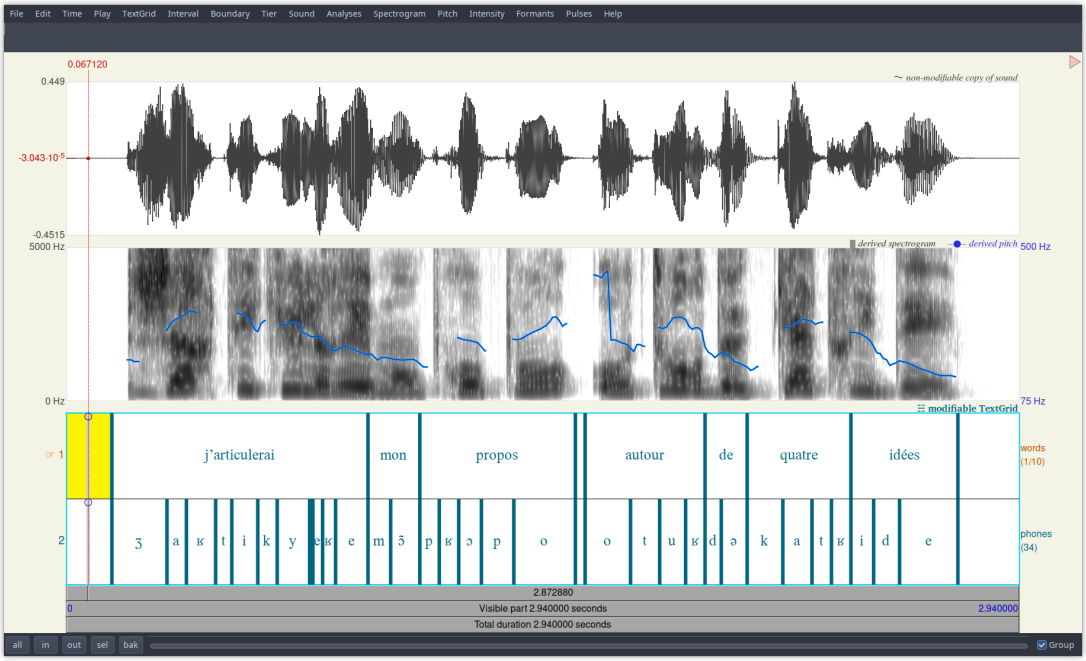
Be aware that you have to use the phone set of the respective dictionary for manual modifications. We run now the validation command again:

In [ ]:
```
! mfa validate --clean ~/Workspace/corpora/french_data/ ~/Workspace/cor
```

As you can see, this time we have no OOV error. We can now align our datasets again:

In [ ]:
```
! mfa align --clean ~/Workspace/corpora/french_data/ ~/Workspace/corpor
```

If we open the new alignment of the file in question, we see that we have now the phone alignments for the once missing word.



Before running the aligner on own or external data, it is best to run the validate command first, in order to find any errors in the corpus (missing files, entries in the dictionary). If it is necessary, the dictionary can always be extended manually. But in case you have multiple words missing, creating a dictionary using a g2p model may be preferred.