Introduction

This is a documentation for 3 scripts whose combined purpose is to reduce the manual workload in text-to-speech alignment when some words are not found in the dictionary being used. Here, I'll refer to the Montreal Forced Aligner (MFA) but you may use other similar aligners relying on .lab and .wav files to run. The individual use of each script is laid out below:

1. GenerateLab.py serves to create .lab files from .wav files and a .docx file. The .lab files can then be used with the .wav files by running the MFA.
2. GenerateDict.py creates a French grapheme-to-phoneme (G2P) dictionary from the out-of-vocabulary (OOV) words following an attempt to use the MFA. It is based on the pronunciation of Basaa speakers. Note that this can be replaced by another program generating a G2P dictionary for another variety of French or for another language.
3. MergeDict.py combines multiple G2P dictionaries to get a broader dictionary which can then be used by the MFA.

I've written these as a Research Assistant under Dr Fatima Hamlaoui at the University of Toronto from June to August 2018.

As of 31 August 2018, the 'front' part of some French words containing apostrophes (such as qu', c', m' etc) are classified by the MFA as OOV even though they are in the dictionary being used. This prevents the generation of TextGrid files but hopefully, this situation will be resolved.

The code can be found at: https://github.com/mchanchee/automate-alignment

Should you need any clarification, you can contact me at: matthieuchanchee@gmail.com

Requirements

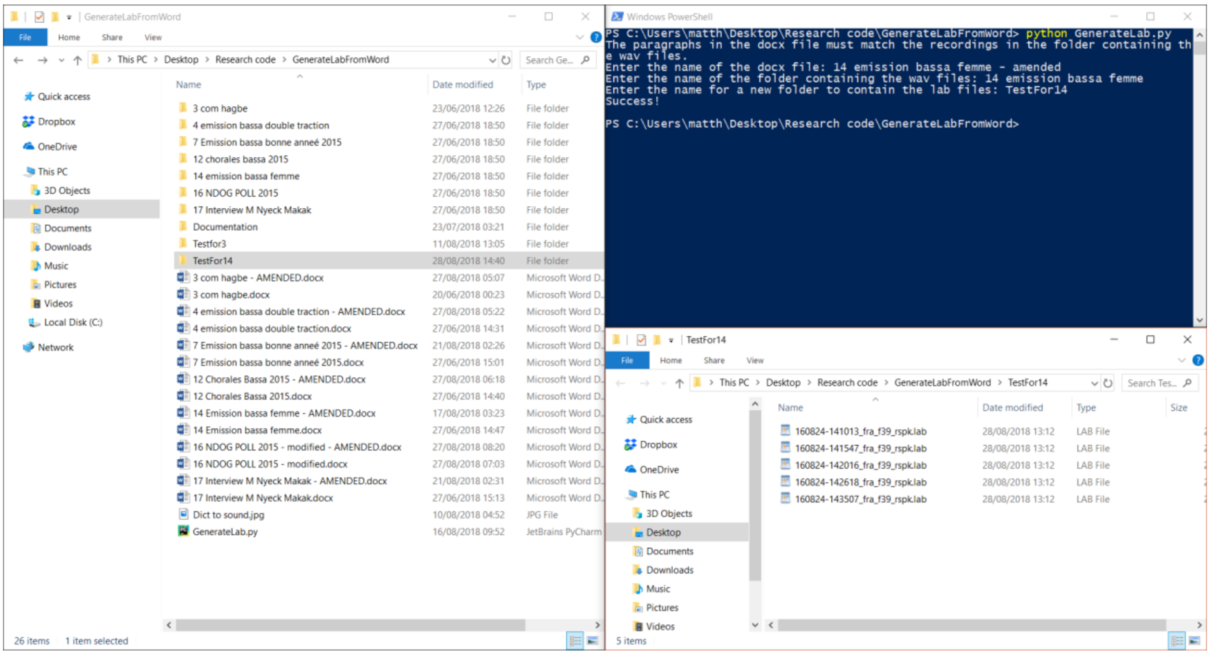
You must have Python 3 installed for all three scripts.

1. GenerateLab.py:
   - The .docx file, the folder containing the .wav files, and GenerateLab.py must be in the same folder.
   - The paragraphs in the .docx file must correspond to the .wav files.
2. GenerateDict.py:
   - The file containing the list of OOVs, the file containing the original G2P dictionary, and GenerateDict.py must be in the same folder.
3. MergeDict.py:
   - The G2P dictionaries to be merged must be in the same folder as MergeDict.py.

Use

1. GeneratLab.py:

   a) Open your cmd/PowerShell/Terminal in the directory containing both the .docx file and the folder consisting of the .wav files. Run GenerateLab.py and follow the steps. A new folder containing the .lab files corresponding to the paragraphs in your .docx file and bearing the same names as the corresponding .wav recordings should be created.



This PC > Desktop > Research code > GenerateLabFromWord

Name	Date modified	Type
3 com hagbe	23/06/2018 12:26	File folder
4 emission bassa double traction	27/06/2018 18:50	File folder
7 Emission bassa bonne anneé 2015	27/06/2018 18:50	File folder
12 chorales bassa 2015	27/06/2018 18:50	File folder
14 emission bassa femme	27/06/2018 18:50	File folder
16 NDOG POLL 2015	27/06/2018 18:50	File folder
17 Interview M Nyeck Makak	27/06/2018 18:50	File folder
Documentation	23/07/2018 03:21	File folder
Testfor3	11/08/2018 13:05	File folder
TestFor14	28/08/2018 14:40	File folder
3 com hagbe - AMENDED.docx	27/08/2018 05:07	Microsoft Word D...
3 com hagbe.docx	20/06/2018 00:23	Microsoft Word D...
4 emission bassa double traction - AMENDED.docx	27/08/2018 05:22	Microsoft Word D...
4 emission bassa double traction.docx	27/06/2018 14:31	Microsoft Word D...
7 Emission bassa bonne anneé 2015 - AMENDED.docx	21/08/2018 02:26	Microsoft Word D...
7 Emission bassa bonne anneé 2015.docx	27/06/2018 15:01	Microsoft Word D...
12 Chorales Bassa 2015 - AMENDED.docx	27/08/2018 06:18	Microsoft Word D...
12 Chorales Bassa 2015.docx	27/06/2018 14:40	Microsoft Word D...
14 Emission bassa femme - AMENDED.docx	17/08/2018 03:23	Microsoft Word D...
14 Emission bassa femme.docx	27/06/2018 14:47	Microsoft Word D...
16 NDOG POLL 2015 - modified - AMENDED.docx	27/08/2018 08:20	Microsoft Word D...
16 NDOG POLL 2015 - modified.docx	27/08/2018 07:03	Microsoft Word D...
17 Interview M Nyeck Makak - AMENDED.docx	21/08/2018 02:31	Microsoft Word D...
17 Interview M Nyeck Makak.docx	27/06/2018 15:13	Microsoft Word D...
Dict to sound.jpg	10/08/2018 04:52	JPG File
GenerateLab.py	16/08/2018 09:52	JetBrains PyCharm

26 items 1 item selected

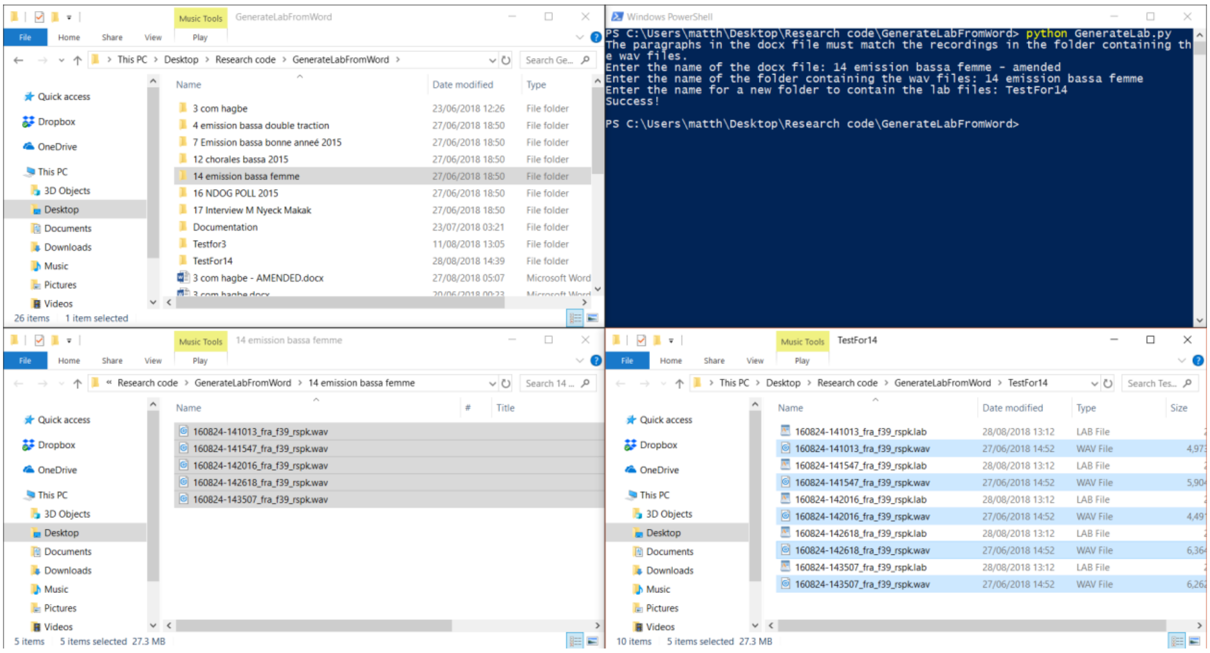
```
PS C:\Users\matth\Desktop\Research code\GenerateLabFromWord> python GenerateLab.py
The paragraphs in the docx file must match the recordings in the folder containing th
e wav files.
Enter the name of the docx file: 14 emission bassa femme - amended
Enter the name of the folder containing the wav files: 14 emission bassa femme
Enter the name for a new folder to contain the lab files: TestFor14
Success!

PS C:\Users\matth\Desktop\Research code\GenerateLabFromWord>
```

This PC > Desktop > Research code > GenerateLabFromWord > TestFor14

Name	Date modified	Type	Size
160824-141013_fra_f39_rspk.lab	28/08/2018 13:12	LAB File	
160824-141547_fra_f39_rspk.lab	28/08/2018 13:12	LAB File	
160824-142016_fra_f39_rspk.lab	28/08/2018 13:12	LAB File	
160824-142618_fra_f39_rspk.lab	28/08/2018 13:12	LAB File	
160824-143507_fra_f39_rspk.lab	28/08/2018 13:12	LAB File	

5 items

   b) Copy and paste the .wav files into the newly created folder.



This PC > Desktop > Research code > GenerateLabFromWord >

Name	Date modified	Type
3 com hagbe	23/06/2018 12:26	File folder
4 emission bassa double traction	27/06/2018 18:50	File folder
7 Emission bassa bonne anneé 2015	27/06/2018 18:50	File folder
12 chorales bassa 2015	27/06/2018 18:50	File folder
14 emission bassa femme	27/06/2018 18:50	File folder
16 NDOG POLL 2015	27/06/2018 18:50	File folder
17 Interview M Nyeck Makak	27/06/2018 18:50	File folder
Documentation	23/07/2018 03:21	File folder
Testfor3	11/08/2018 13:05	File folder
TestFor14	28/08/2018 14:39	File folder
3 com hagbe - AMENDED.docx	27/08/2018 05:07	Microsoft Word
3 com hagbe.docx	20/06/2018 00:23	Microsoft Word

26 items 1 item selected

```
PS C:\Users\matth\Desktop\Research code\GenerateLabFromWord> python GenerateLab.py
The paragraphs in the docx file must match the recordings in the folder containing th
e wav files.
Enter the name of the docx file: 14 emission bassa femme - amended
Enter the name of the folder containing the wav files: 14 emission bassa femme
Enter the name for a new folder to contain the lab files: TestFor14
Success!

PS C:\Users\matth\Desktop\Research code\GenerateLabFromWord>
```

« Research code > GenerateLabFromWord > 14 emission bassa femme

Name	#	Title
160824-141013_fra_f39_rspk.wav		
160824-141547_fra_f39_rspk.wav		
160824-142016_fra_f39_rspk.wav		
160824-142618_fra_f39_rspk.wav		
160824-143507_fra_f39_rspk.wav		

5 items 5 items selected 27.3 MB

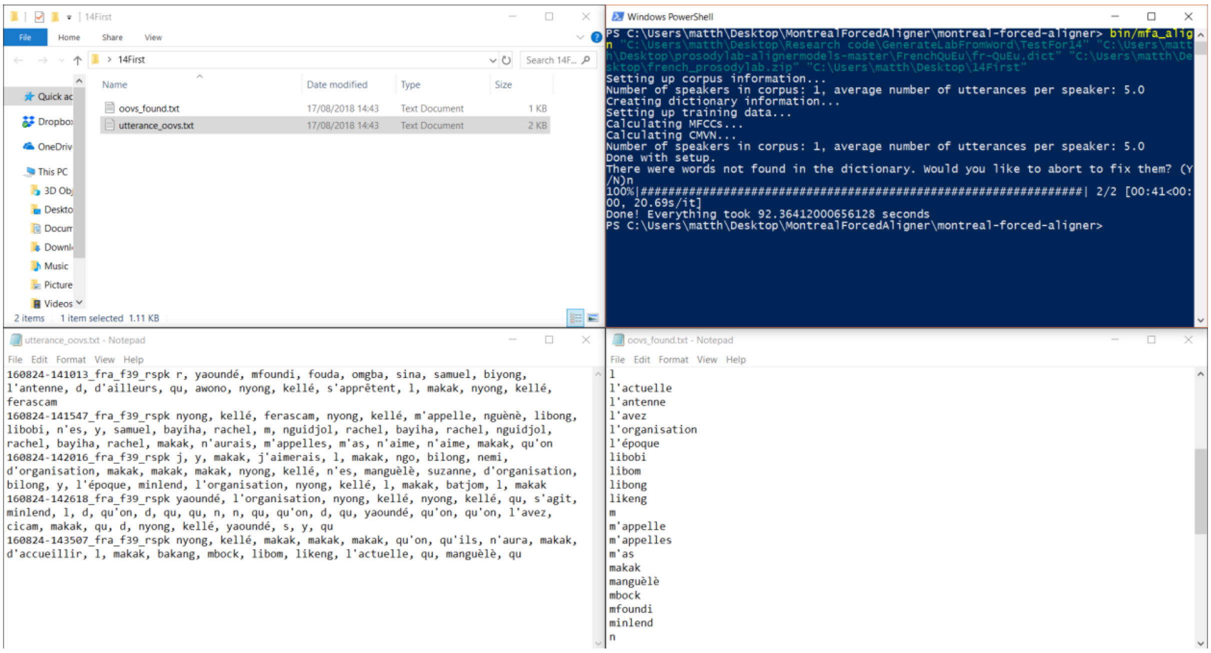
This PC > Desktop > Research code > GenerateLabFromWord > TestFor14

Name	Date modified	Type	Size
160824-141013_fra_f39_rspk.lab	28/08/2018 13:12	LAB File	
160824-141013_fra_f39_rspk.wav	27/06/2018 14:52	WAV File	4,97
160824-141547_fra_f39_rspk.lab	28/08/2018 13:12	LAB File	
160824-141547_fra_f39_rspk.wav	27/06/2018 14:52	WAV File	5,90
160824-142016_fra_f39_rspk.lab	28/08/2018 13:12	LAB File	
160824-142016_fra_f39_rspk.wav	27/06/2018 14:52	WAV File	4,49
160824-142618_fra_f39_rspk.lab	28/08/2018 13:12	LAB File	
160824-142618_fra_f39_rspk.wav	27/06/2018 14:52	WAV File	6,36
160824-143507_fra_f39_rspk.lab	28/08/2018 13:12	LAB File	
160824-143507_fra_f39_rspk.wav	27/06/2018 14:52	WAV File	6,26

10 items 5 items selected 27.3 MB

2. GenerateDict.py:

a) Run the MFA. Here, the Prosodylab G2P dictionary (https://github.com/prosodylab/prosodylab-alignermodels/tree/master/FrenchQuEu) and acoustic model (https://montreal-forced-aligner.readthedocs.io/en/latest/pretrained_models.html) are being used. The folder 14First is created by the MFA.



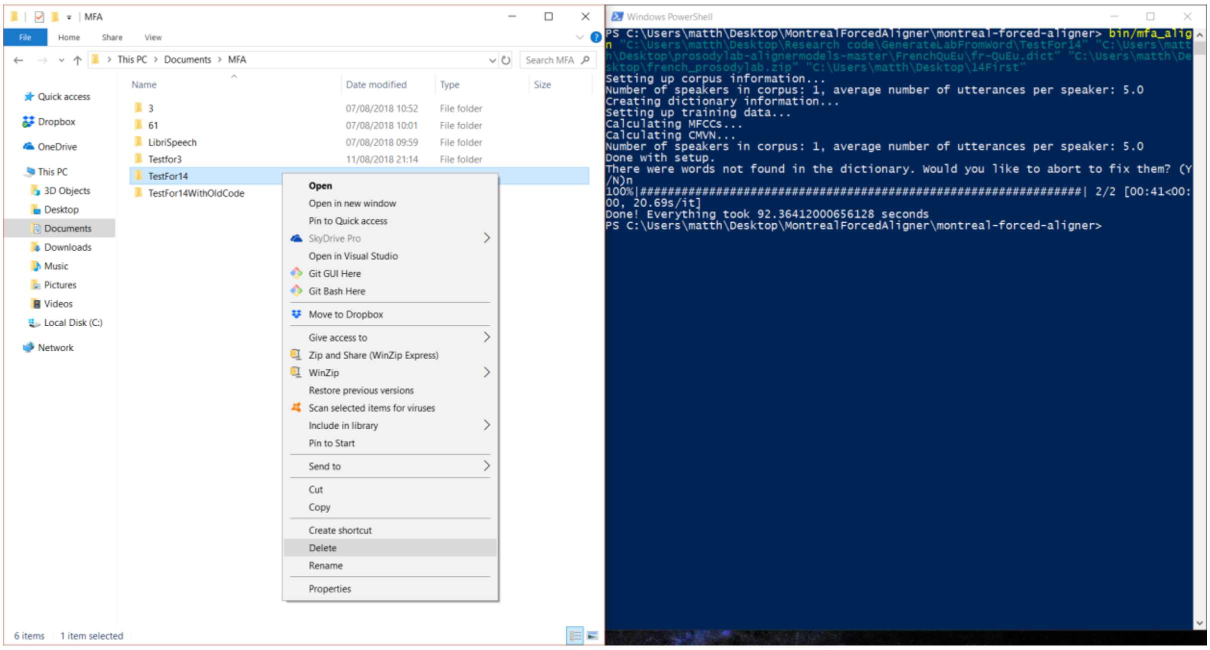
Name	Date modified	Type	Size
oovs_found.txt	17/08/2018 14:43	Text Document	1 KB
utterance_oovs.txt	17/08/2018 14:43	Text Document	2 KB

```
PS C:\Users\matth\Desktop\MontrealForcedAligner\montreal-forced-aligner> bin/mfa_alig
n "C:\Users\matth\Desktop\Research code\GenerateLabFromWord\TestFor14" "C:\Users\matt
h\Desktop\prosodylab-alignermodels-master\FrenchQuEu\fr-QuEu.dict" "C:\Users\matth\De
sktop\french_prosodylab.zip" "C:\Users\matth\Desktop\14First"
Setting up corpus information...
Number of speakers in corpus: 1, average number of utterances per speaker: 5.0
Creating dictionary information...
Setting up training data...
Calculating MFCCs...
Calculating CMVN...
Number of speakers in corpus: 1, average number of utterances per speaker: 5.0
Done with setup.
There were words not found in the dictionary. Would you like to abort to fix them? (Y
/N)n
100%|################################################################| 2/2 [00:41<00:
00, 20.69s/it]
Done! Everything took 92.36412000656128 seconds
PS C:\Users\matth\Desktop\MontrealForcedAligner\montreal-forced-aligner>
```

```
160824-141013_fra_f39_rspk r, yaoundé, mfoundi, fouda, omgba, sina, samuel, biyong,
l'antenne, d, d'ailleurs, qu, awono, nyong, kellé, s'apprêtent, l, makak, nyong, kellé,
ferascam
160824-141547_fra_f39_rspk nyong, kellé, ferascam, nyong, kellé, m'appelle, nguènè, libong,
libobi, n'es, y, samuel, bayiha, rachel, m, nguidjol, rachel, bayiha, rachel, nguidjol,
rachel, bayiha, rachel, makak, n'aurais, m'appelles, m'as, n'aime, n'aime, makak, qu'on
160824-142016_fra_f39_rspk j, y, makak, j'aimerais, l, makak, ngo, bilong, nemi,
d'organisation, makak, makak, makak, nyong, kellé, n'es, manguèlè, suzanne, d'organisation,
bilong, y, l'époque, minlend, l'organisation, nyong, kellé, l, makak, batjom, l, makak
160824-142618_fra_f39_rspk yaoundé, l'organisation, nyong, kellé, nyong, kellé, qu, s'agit,
minlend, l, d, qu'on, d, qu, qu, n, n, qu, qu'on, d, qu, yaoundé, qu'on, qu'on, l'avez,
cicam, makak, qu, d, nyong, kellé, yaoundé, s, y, qu
160824-143507_fra_f39_rspk nyong, kellé, makak, makak, makak, qu'on, qu'ils, n'aura, makak,
d'accueillir, l, makak, bakang, mbock, libom, likeng, l'actuelle, qu, manguèlè, qu
```

```
l
l'actuelle
l'antenne
l'avez
l'organisation
l'époque
libobi
libom
libong
likeng
m
m'appelle
m'appelles
m'as
makak
manguèlè
mbock
mfoundi
minlend
n
```

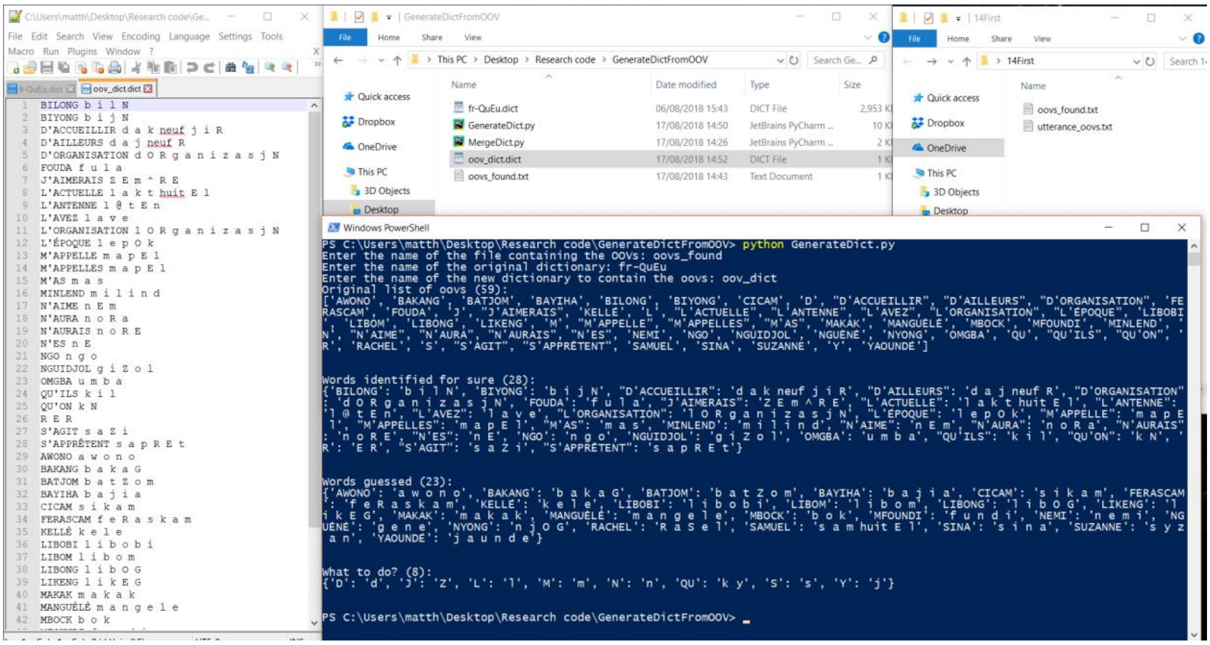
b) Delete the folder containing the logs generated by the MFA. This is because we'll use the name 'TestFor14' again later.



Name	Date modified	Type	Size
3	07/08/2018 10:52	File folder	
61	07/08/2018 10:01	File folder	
LibriSpeech	07/08/2018 09:59	File folder	
Testfor3	11/08/2018 21:14	File folder	
TestFor14			
TestFor14WithOldCode			

```
PS C:\Users\matth\Desktop\MontrealForcedAligner\montreal-forced-aligner> bin/mfa_alig
n "C:\Users\matth\Desktop\Research code\GenerateLabFromWord\TestFor14" "C:\Users\matt
h\Desktop\prosodylab-alignermodels-master\FrenchQuEu\fr-QuEu.dict" "C:\Users\matth\De
sktop\french_prosodylab.zip" "C:\Users\matth\Desktop\14First"
Setting up corpus information...
Number of speakers in corpus: 1, average number of utterances per speaker: 5.0
Creating dictionary information...
Setting up training data...
Calculating MFCCs...
Calculating CMVN...
Number of speakers in corpus: 1, average number of utterances per speaker: 5.0
Done with setup.
There were words not found in the dictionary. Would you like to abort to fix them? (Y
/N)n
100%|################################################################| 2/2 [00:41<00:
00, 20.69s/it]
Done! Everything took 92.36412000656128 seconds
PS C:\Users\matth\Desktop\MontrealForcedAligner\montreal-forced-aligner>
```

c) Copy and paste our original dictionary (here, fr-QuEu) and the file containing the list of OOVs (here, oovs_found) into the folder containing GenerateDict.py. Then, run the latter to create a G2P dictionary for these OOVs (here, oov_dict).



```
BILONG b i l N
BIYONG b i j N
D'ACCUEILLIR d a k neuf j i R
D'AILLEURS d a j neuf R
D'ORGANISATION d O R g a n i z a s j N
FOUDA f u l a
J'AIMERAIS Z E m ^ R E
L'ACTUELLE l a k t huit E l
L'ANTENNE l @ t E n
L'AVEZ l a v e
L'ORGANISATION l O R g a n i z a s j N
L'ÉPOQUE l e p O k
M'APPELLE m a p E l
M'APPELLES m a p E l
M'AS m a s
MINLEND m i l i n d
N'AIME n E m
N'AURA n o R a
N'AURAIS n o R E
N'ES n E
NGO n g o
NGUIDJOL g i Z o l
OMGBA u m b a
QU'ILS k i l
QU'ON k N
R E R
S'AGIT s a Z i
S'APPRÊTENT s a p R E t
AWONO a w o n o
BAKANG b a k a G
BATJOM b a t Z o m
BAYIHA b a j i a
CICAM s i k a m
FERASCAM f e R a s k a m
KELLÉ k e l e
LIBOBI l i b o b i
LIBOM l i b o m
LIBONG l i b O G
LIKENG l i k E G
MAKAK m a k a k
MANGUÈLÈ m a n g e l e
MBOCK b o k
```

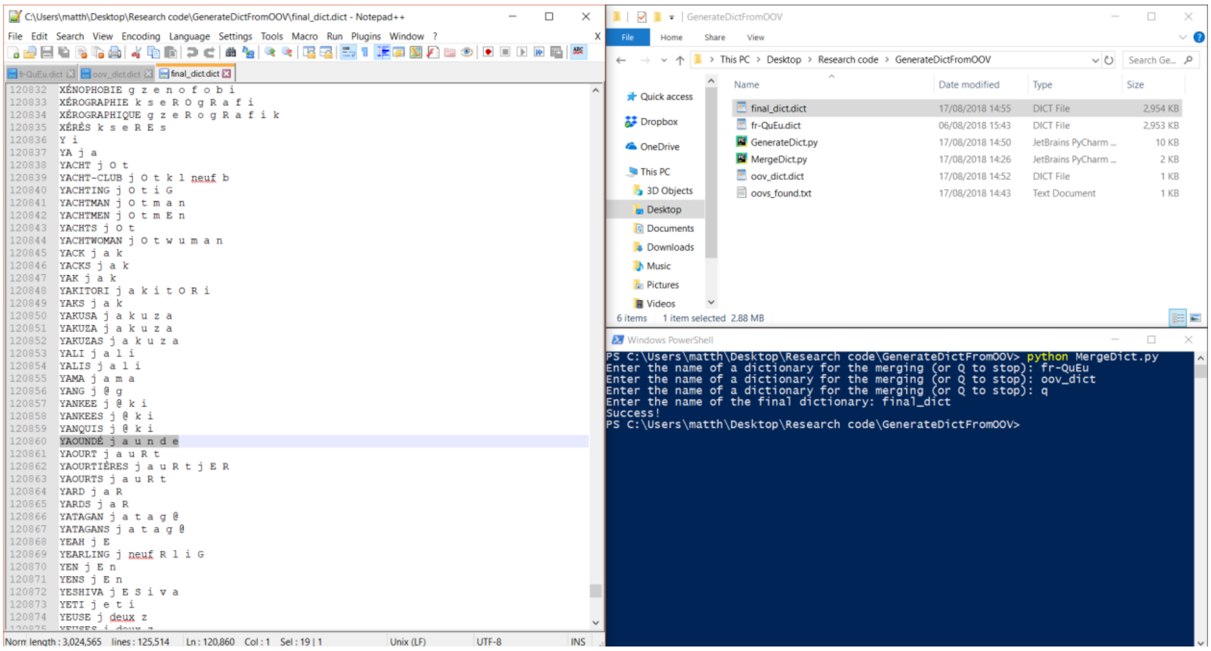
Name	Date modified	Type	Size
fr-QuEu.dict	06/08/2018 15:43	DICT File	2,953 K
GenerateDict.py	17/08/2018 14:50	JetBrains PyCharm ...	10 K
MergeDict.py	17/08/2018 14:26	JetBrains PyCharm ...	2 K
oov_dict.dict	17/08/2018 14:52	DICT File	1 K
oovs_found.txt	17/08/2018 14:43	Text Document	1 K

Name
oovs_found.txt
utterance_oovs.txt

```
PS C:\Users\matth\Desktop\Research code\GenerateDictFromOOV> python GenerateDict.py
Enter the name of the file containing the OOVs: oovs_found
Enter the name of the original dictionary: fr-QuEu
Enter the name of the new dictionary to contain the oovs: oov_dict
Original list of oovs (59):
['AWONO', 'BAKANG', 'BATJOM', 'BAYIHA', 'BILONG', 'BIYONG', 'CICAM', 'D', "D'ACCUEILLIR", "D'AILLEURS", "D'ORGANISATION", 'FE
RASCAM', 'FOUDA', 'J', "J'AIMERAIS", 'KELLÉ', 'L', "L'ACTUELLE", "L'ANTENNE", "L'AVEZ", "L'ORGANISATION", "L'ÉPOQUE", 'LIBOBI
', 'LIBOM', 'LIBONG', 'LIKENG', 'M', "M'APPELLE", "M'APPELLES", "M'AS", 'MAKAK', 'MANGUÈLÈ', 'MBOCK', 'MFOUNDI', 'MINLEND', '
N', "N'AIME", "N'AURA", "N'AURAIS", "N'ES", 'NEMI', 'NGO', 'NGUIDJOL', 'NGUÈNÈ', 'NYONG', 'OMGBA', 'QU', "QU'ILS", "QU'ON", '
R', 'RACHEL', 'S', "S'AGIT", "S'APPRÊTENT", 'SAMUEL', 'SINA', 'SUZANNE', 'Y', 'YAOUNDÉ']


Words identified for sure (28):
{'BILONG': 'b i l N', 'BIYONG': 'b i j N', "D'ACCUEILLIR": 'd a k neuf j i R', "D'AILLEURS": 'd a j neuf R', "D'ORGANISATION"
: 'd O R g a n i z a s j N', 'FOUDA': 'f u l a', "J'AIMERAIS": 'Z E m ^ R E', "L'ACTUELLE": 'l a k t huit E l', "L'ANTENNE":
'l @ t E n', "L'AVEZ": 'l a v e', "L'ORGANISATION": 'l O R g a n i z a s j N', "L'ÉPOQUE": 'l e p O k', "M'APPELLE": 'm a p E
 l', "M'APPELLES": 'm a p E l', "M'AS": 'm a s', 'MINLEND': 'm i l i n d', "N'AIME": 'n E m', "N'AURA": 'n o R a', "N'AURAIS"
: 'n o R E', "N'ES": 'n E', 'NGO': 'n g o', 'NGUIDJOL': 'g i Z o l', 'OMGBA': 'u m b a', "QU'ILS": 'k i l', "QU'ON": 'k N', '
R': 'E R', "S'AGIT": 's a Z i', "S'APPRÊTENT": 's a p R E t'}


Words guessed (23):
{'AWONO': 'a w o n o', 'BAKANG': 'b a k a G', 'BATJOM': 'b a t Z o m', 'BAYIHA': 'b a j i a', 'CICAM': 's i k a m', 'FERASCAM
': 'f e R a s k a m', 'KELLÉ': 'k e l e', 'LIBOBI': 'l i b o b i', 'LIBOM': 'l i b o m', 'LIBONG': 'l i b O G', 'LIKENG': 'l
i k E G', 'MAKAK': 'm a k a k', 'MANGUÈLÈ': 'm a n g e l e', 'MBOCK': 'b o k', 'MFOUNDI': 'f u n d i', 'NEMI': 'n e m i', 'NG
UÈNÈ': 'g e n e', 'NYONG': 'n j O G', 'RACHEL': 'R a S e l', 'SAMUEL': 's a m huit E l', 'SINA': 's i n a', 'SUZANNE': 's y z
 a n', 'YAOUNDÉ': 'j a u n d e'}


What to do? (8):
{'D': 'd', 'J': 'Z', 'L': 'l', 'M': 'm', 'N': 'n', 'QU': 'k y', 'S': 's', 'Y': 'j'}


PS C:\Users\matth\Desktop\Research code\GenerateDictFromOOV> _
```

3. MergeDict.py:

a) Run MergeDict.py to combine our original dictionary and our OOV dictionary. This gives final_dict here.



C:\Users\matth\Desktop\Research code\GenerateDictFromOOV\final_dict.dict - Notepad++

```
XÉNOPHOBIE g z e n o f o b i
XÉROGRAPHIE k s e R O g R a f i
XÉROGRAPHIQUE g z e R o g R a f i k
XÉRÈS k s e R E s
Y i
YA j a
YACHT j O t
YACHT-CLUB j O t k l neuf b
YACHTING j O t i G
YACHTMAN j O t m a n
YACHTMEN j O t m E n
YACHTS j O t
YACHTWOMAN j O t w u m a n
YACK j a k
YACKS j a k
YAK j a k
YAKITORI j a k i t O R i
YAKS j a k
YAKUSA j a k u z a
YAKUZA j a k u z a
YAKUZAS j a k u z a
YALI j a l i
YALIS j a l i
YAMA j a m a
YANG j @ g
YANKEE j @ k i
YANKEES j @ k i
YANQUIS j @ k i
YAOUNDÉ j a u n d e
YAOURT j a u R t
YAOURTIÈRES j a u R t j E R
YAOURTS j a u R t
YARD j a R
YARDS j a R
YATAGAN j a t a g @
YATAGANS j a t a g @
YEAH j E
YEARLING j neuf R l i G
YEN j E n
YENS j E n
YESHIVA j E S i v a
YETI j e t i
YEUSE j deux z
```

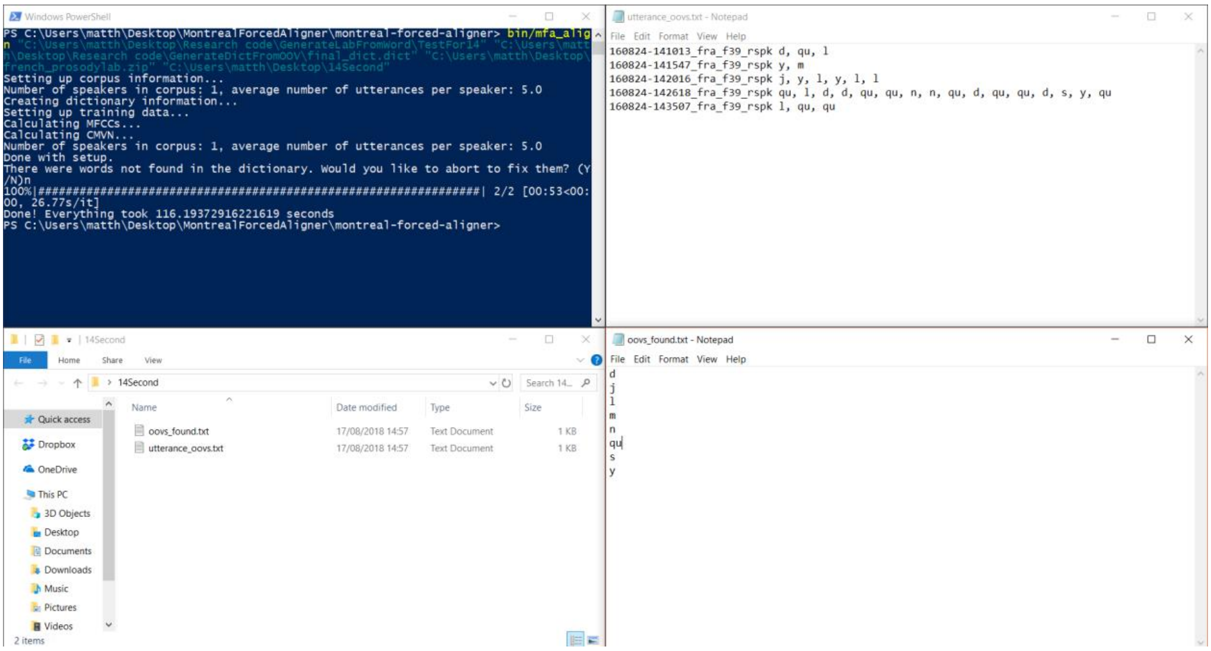
Norm length : 3,024,565 lines : 125,514 Ln : 120,860 Col : 1 Sel : 19 | 1 Unix (LF) UTF-8 INS

Name	Date modified	Type	Size
final_dict.dict	17/08/2018 14:55	DICT File	2,954 KB
fr-QuEu.dict	06/08/2018 15:43	DICT File	2,953 KB
GenerateDict.py	17/08/2018 14:50	JetBrains PyCharm ...	10 KB
MergeDict.py	17/08/2018 14:26	JetBrains PyCharm ...	2 KB
oov_dict.dict	17/08/2018 14:52	DICT File	1 KB
oovs_found.txt	17/08/2018 14:43	Text Document	1 KB

6 items 1 item selected 2.88 MB

```
PS C:\Users\matth\Desktop\Research code\GenerateDictFromOOV> python MergeDict.py
Enter the name of a dictionary for the merging (or Q to stop): fr-QuEu
Enter the name of a dictionary for the merging (or Q to stop): oov_dict
Enter the name of a dictionary for the merging (or Q to stop): q
Enter the name of the final dictionary: final_dict
Success!
PS C:\Users\matth\Desktop\Research code\GenerateDictFromOOV>
```

b) Finally, run the MFA with our newly generated dictionary.



```
PS C:\Users\matth\Desktop\MontrealForcedAligner\montreal-forced-aligner> bin/mfa_alig
n "C:\Users\matth\Desktop\Research code\GenerateLabFromWord\TestFor14" "C:\Users\matt
h\Desktop\Research code\GenerateDictFromOOV\final_dict.dict" "C:\Users\matth\Desktop\
french_prosodylab.zip" "C:\Users\matth\Desktop\14Second"
Setting up corpus information...
Number of speakers in corpus: 1, average number of utterances per speaker: 5.0
Creating dictionary information...
Setting up training data...
Calculating MFCCs...
Calculating CMVN...
Number of speakers in corpus: 1, average number of utterances per speaker: 5.0
Done with setup.
There were words not found in the dictionary. Would you like to abort to fix them? (Y
/N)n
100%|###################################################################| 2/2 [00:53<00:
00, 26.77s/it]
Done! Everything took 116.19372916221619 seconds
PS C:\Users\matth\Desktop\MontrealForcedAligner\montreal-forced-aligner>
```

```
160824-141013_fra_f39_rspk d, qu, l
160824-141547_fra_f39_rspk y, m
160824-142016_fra_f39_rspk j, y, l, y, l, l
160824-142618_fra_f39_rspk qu, l, d, d, qu, qu, n, n, qu, d, qu, qu, d, s, y, qu
160824-143507_fra_f39_rspk l, qu, qu
```

Name	Date modified	Type	Size
oovs_found.txt	17/08/2018 14:57	Text Document	1 KB
utterance_oovs.txt	17/08/2018 14:57	Text Document	1 KB

```
d
j
l
m
n
qu
s
y
```

Notes

On all 3 scripts

In the following cases, an error message will be displayed and no file will be generated:

- A file or folder (e.g. the .docx file and the folder containing .wav files on step 1a) whose name was input does not exist.
- Any of the following characters was used to name a new file or folder:

  \, /, :, *, ?, ", <, >, |
- The name you entered for a new file or folder is already used.

On GenerateLab.py

An error message will be displayed and no file will be generated if number of paragraphs in the .docx file is not equal to the number of .wav files.

Some people end their .docx files by pressing several times on the Enter/Return key, thus creating ‘empty’ paragraphs. This program ignores empty paragraphs I’ve been able to predict but should other types of empty paragraphs show up, the code will have to be updated. As of 31 August 2018, an empty paragraph is defined as any of:

- Nothing (empty string)
- 1 to 9 spaces (‘ ’, ‘ ’, …, ‘   ’)