Automate Alignment Documentation V2

Table of Contents

1 Introduction

This is a documentation for all the work that was done for the generation of Textgrid files from .wav and .txt files. This includes converting file types, how to use the Montreal Forced Aligner, as well as cleaning the raw data. Here, Montreal Forced Aligner (MFA) will be referred to but similar aligners relying on .lab and .wav files like the Prosodylab Forced Aligner may be used.

The scripts could be found at https://github.com/Frank-XNS/AutomateAlignmentV2.
This documentation was written by Sun XIaonong, who was a research assistant with Dr Fatima Hamlaoui from 5th September to 22nd February. If you have any questions, feel free to reach out to me at frankxns.sun@mail.utoronto.ca.

2 The Montreal Forced Aligner

2.1 Introduction

The gist of the Montreal Forced Aligner (MFA) is that it is a Forced Alignment script that takes spoken dialogue (in the form of .wav files) and its annotations (in the form of .lab files) and produce a file (in the form of .textgrid) that matches each phoneme of each word specified in the annotation to the corresponding time that that phoneme is spoken in the dialogue. The .textgrid file will be used alongside the .wav file in a Phonetic Analysis program like PRAAT for linguistic studies.

In order to achieve this, the Montreal Forced Aligner requires three things

1. The folder containing the .lab and .wav file together, or the corpus dictionary.
2. The pronunciation dictionary. This maps the Grapheme to its corresponding phoneme. This could be custom or downloaded online. The online pronunciation dictionary primary used is the French Prosodylab dictionary, which can be found on https://raw.githubusercontent.com/prosodylab/prosodylab.dictionaries/master/fr.dict.
   However the pronunciation dictionary downloaded online is not complete, and OOVs can be used to expand the dictionary. See section 2.2 for more details.
3. The acoustic model. See Section 2.3.2 below for more details

In most cases, you will have .docx files but not .lab files, there are scripts provided to clean the data and also convert .docx to .lab files, please refer to section 3 for details on the scripts. For downloading, requirements and additional information for MFA, please visit the MFA website at https://montreal-forced-aligner.readthedocs.io/en/latest/index.html

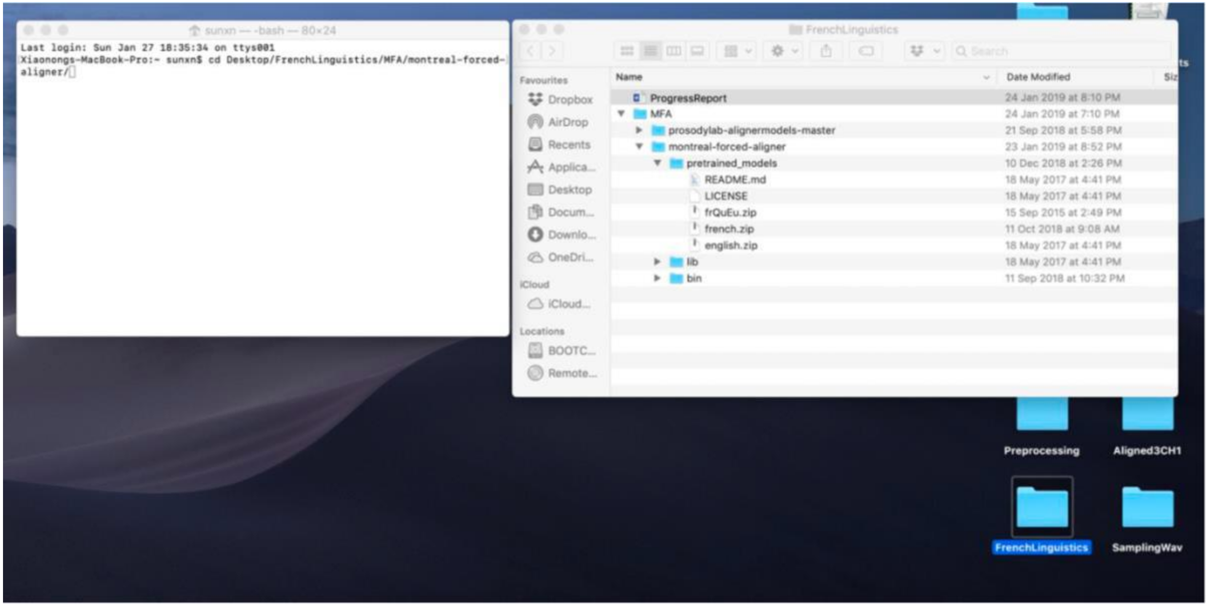
2.2 Out-Of-Vocabulary-Words (OOVs)

OOVs are words that are not currently in the pronunciation dictionary, and will be outputted by the MFA as a separate .txt file in the output path. There exists a script that generates a phonetic pronunciation of the words in the OOV file and also a script that combines the OOV file with a pre-existing dictionary to form a more complete dictionary. See section 3.3 and 3.4 for more information. As well, a helper function exists that combines several OOVs into one without duplicates such that it is cleaner and easier to read. See section 4.2 for more information.

2.3 Running the Montreal Forced Aligner

2.3.1 Preparations

Before running the program, navigate into the MFA folder first in terminal as shown in Fig 1.



```
Last login: Sun Jan 27 18:35:34 on ttys001
Xiaonongs-MacBook-Pro:~ sunxn$ cd Desktop/FrenchLinguistics/MFA/montreal-forced-
aligner/
```

Name	Date Modified
ProgressReport	24 Jan 2019 at 8:10 PM
MFA	24 Jan 2019 at 7:10 PM
prosodylab-alignermodels-master	21 Sep 2018 at 5:58 PM
montreal-forced-aligner	23 Jan 2019 at 8:52 PM
pretrained_models	10 Dec 2018 at 2:26 PM
README.md	18 May 2017 at 4:41 PM
LICENSE	18 May 2017 at 4:41 PM
frQuEu.zip	15 Sep 2015 at 2:49 PM
french.zip	11 Oct 2018 at 9:08 AM
english.zip	18 May 2017 at 4:41 PM
lib	18 May 2017 at 4:41 PM
bin	11 Sep 2018 at 10:32 PM

Fig 1. Navigation in the MFA

2.3.2 To generate acoustic model from pre-existing data

An acoustic model is something that maps a phoneme to a sound and is necessary for the MFA to work. However, the pre-existing models provided on the website does not work. The way to get around that is to generate an acoustic model manually using existing data. To do that, type the following into the command prompt

```
bin/mfa_train_and_align corpus_directory dictionary_path output_file -o path
```

Where the corpus_directory is the corpus dictionary (note: the .wav files have to have the same sampling rate, or else an error will be thrown), the dictionary_path is the path to where the pronunciation dictionary is kept, and the output file is the location of the folder of the output (the Textgrid and OOVs). The -o path is necessary if you want to generate an acoustic model, and a zip file containing the acoustic model is placed in the target folder path (path). If the "-o path" does not exist, .textgrid and OOVs will still be generated, but there will be no acoustic models.

2.3.3 To generate a Textgrid file with an existing acoustic model

If you already have an acoustic model and just want the .textgrid and OOV files, type the following into the command prompt

```
bin/mfa_align corpus_directory dictionary_path acoustic_model_path output_file
```

Where the acoustic_model_path is the path to the acoustic model, and the rest are the same as defined in Section 2.3.2.

Caution
If you use an acoustic model that was generated using pre-existing data, when you run mfa_align, make sure you use the SAME dictionary that was used to generate that acoustic model, or else no Textgrid will be generated. This is also the reason why the acoustic models provided on the website does not work.

3 Automate Alignment

3.1 Introduction

The 3 scripts below has the combined purpose of reducing the manual workload in text-to-speech alignment when some words are not found in the dictionary being used. Starting from raw data (.docx file and a .wav file), the scripts are typically run in the following order:

GenerateLab.py -> GenerateDict.py -> MergeDict.py

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.

3.2 GenerateLab.py

3.2.1 Introduction

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. This scripts also converts numbers found in the paragraphs into words as well as remove the symbols , : . – ‘ ' such that the MFA can recognize them.

3.2.2 Requirements

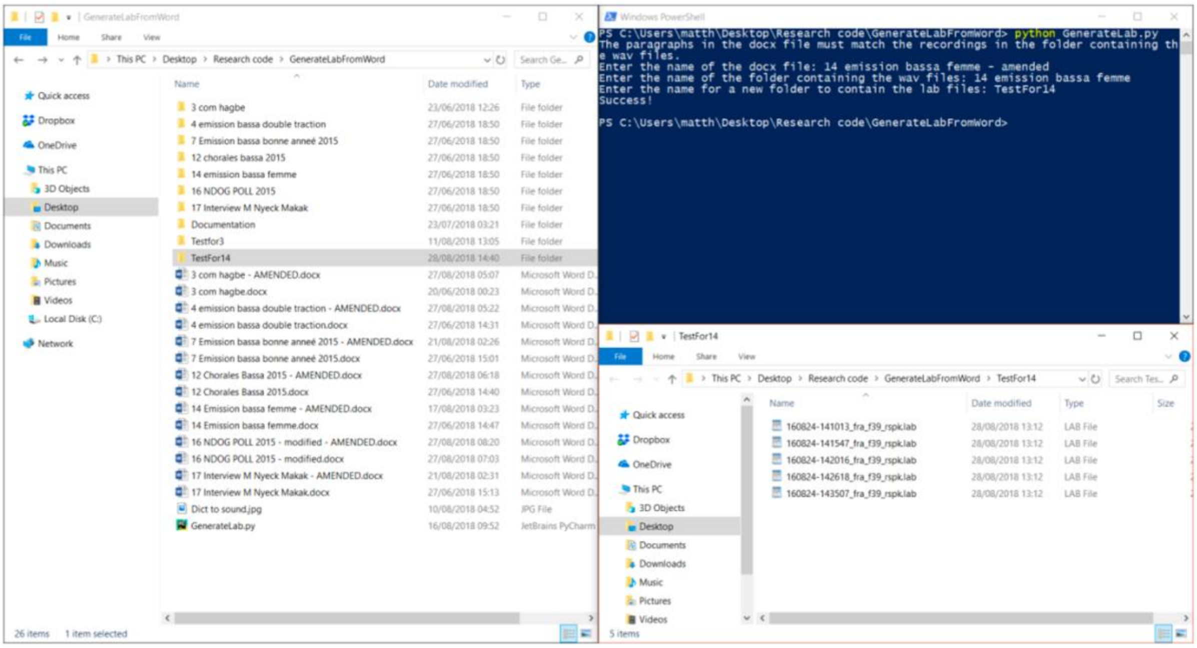
- Python 3 must be installed on your computer
- The .docx file, the folder containing the .wav files, and GenerateLab.py must be in the same folder
- The paragraphs in the .docx file must correspond to the .wav files

Note:
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 March 1st 2019, an empty paragraph is defined as any of:

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

3.2.3 Running

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.



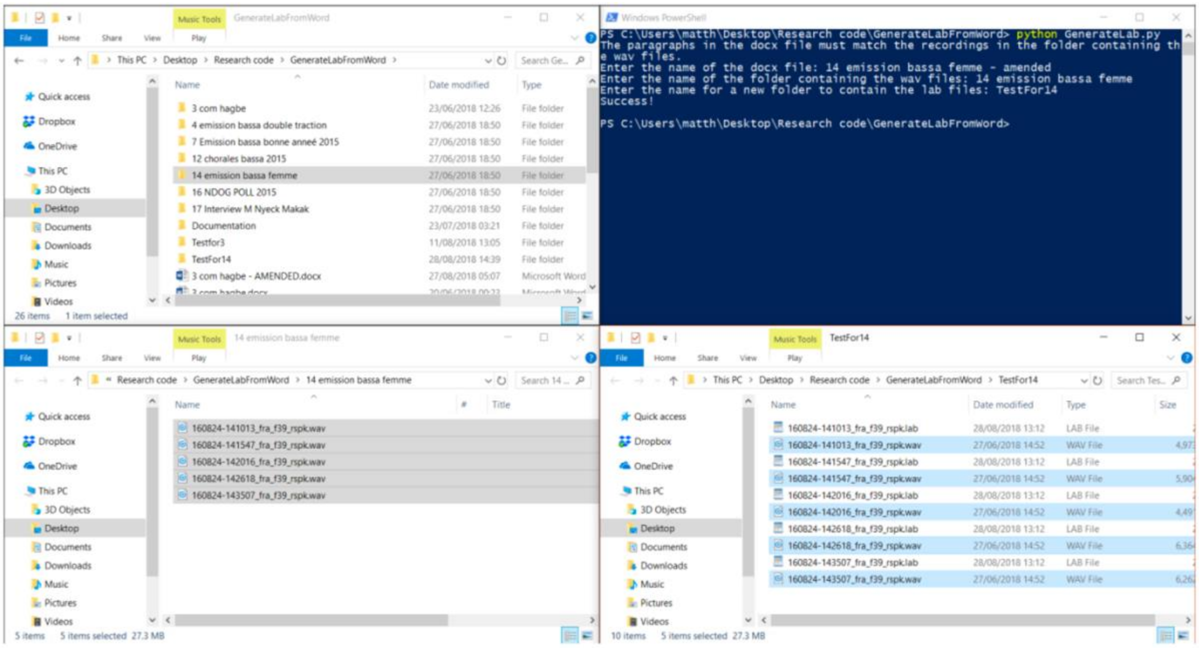
Name	Date modified	Type
3 com hagbe	23/06/2018 12:26	File folder
4 emission bassa double traction	27/06/2018 18:50	File folder
7 Emission bassa bonne anneé 2015	27/06/2018 18:50	File folder
12 chorales bassa 2015	27/06/2018 18:50	File folder
14 emission bassa femme	27/06/2018 18:50	File folder
16 NDOG 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

```
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>
```

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	

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


GenerateLabFromWord
This PC › Desktop › Research code › GenerateLabFromWord ›
Quick access
Dropbox
OneDrive
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
26 items 1 item selected
Windows PowerShell
14 emission bassa femme
« Research code › GenerateLabFromWord › 14 emission bassa femme
5 items 5 items selected 27.3 MB
TestFor14
This PC › Desktop › Research code › GenerateLabFromWord › TestFor14
10 items 5 items selected 27.3 MB

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

```
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>
```

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		

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

In some cases, the script will ask if you want the words pronounced in vigésimales system or regular number[1], the two systems are defined below.

Year	Vigésimales System	Regular Number
1999	dix-neuf cent quatre-vingt-dix-neuf	mille neuf cent quatre-vingt-dix-neuf

The script also asks you if the .wav files are in ascending name. Select true if the order of the naming of the .wav file corresponds to the order that the paragraphs (as annotated in .lab files) are spoken.

3.3 GenerateDict.py

3.3.1 Introduction

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.

[1] See https://www.thoughtco.com/years-in-french-1368976 for additional information

Note: GenerateDict.py converts the following grapheme into the phonemes

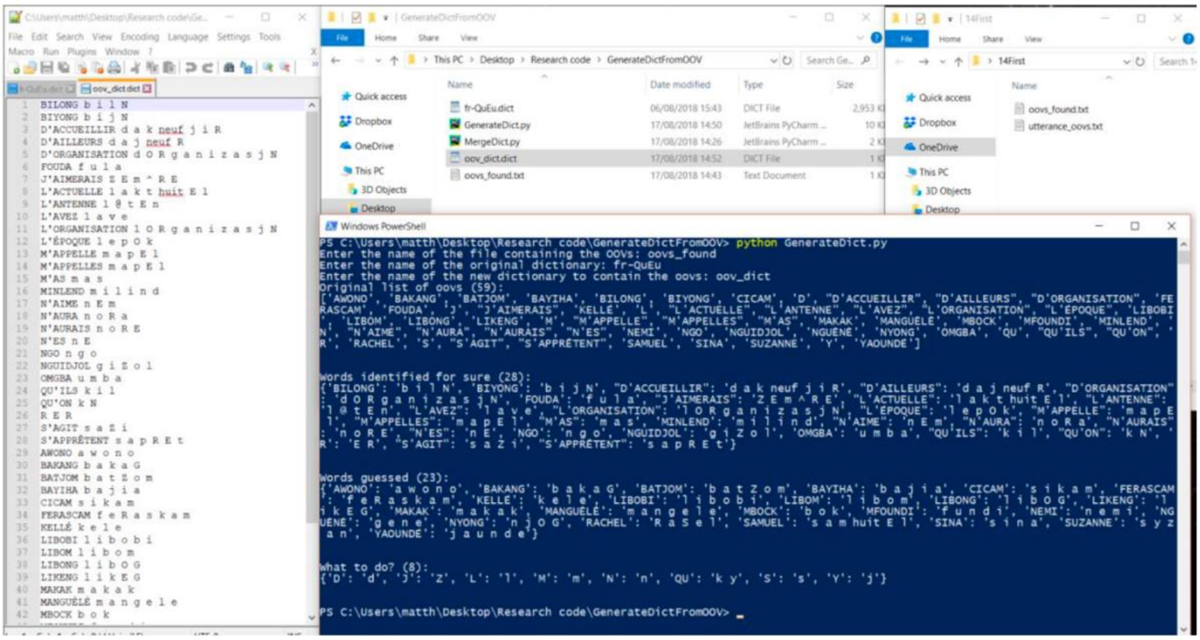
- ô to o
- ï to i
- à to a
- ê to E
- ë depends on the word but usually E
- â -to a

3.3.2 Requirements

- Python3 must be installed on your computer
- The file containing the list of OOVs, the file containing the original G2P dictionary, and GenerateDict.py must be in the same folder

3.3.3 Running

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 and follow the onscreen instructions to create a G2P dictionary for these OOVs (here, oov_dict).


C:\Users\matth\Desktop\Research code\Ge...
File Edit Search View Encoding Language Settings Tools
Macro Run Plugins Window ?
fr-QuEu.dict
oov_dict.dict
GenerateDictFromOOV
File Home Share View
This PC > Desktop > Research code > GenerateDictFromOOV
Search Ge...
Quick access
Dropbox
OneDrive
This PC
3D Objects
Desktop
Name Date modified Type Size
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
14First
14First
oovs_found.txt
utterance_oovs.txt
Windows PowerShell

```
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
MANGUÉLÉ m a n g e l e
MBOCK b o k
```

```
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', 'MANGUÉLÉ', 'MBOCK', 'MFOUNDI', 'MINLEND', '
N', "N'AIME", "N'AURA", "N'AURAIS", "N'ES", 'NEMI', 'NGO', 'NGUIDJOL', 'NGUÉNÉ', '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', 'MANGUÉLÉ': 'm a n g e l e', 'MBOCK': 'b o k', 'MFOUNDI': 'f u n d i', 'NEMI': 'n e m i', 'NG
UÉNÉ': '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> _
```

3.4 MergeDict.py

3.4.1 Introduction

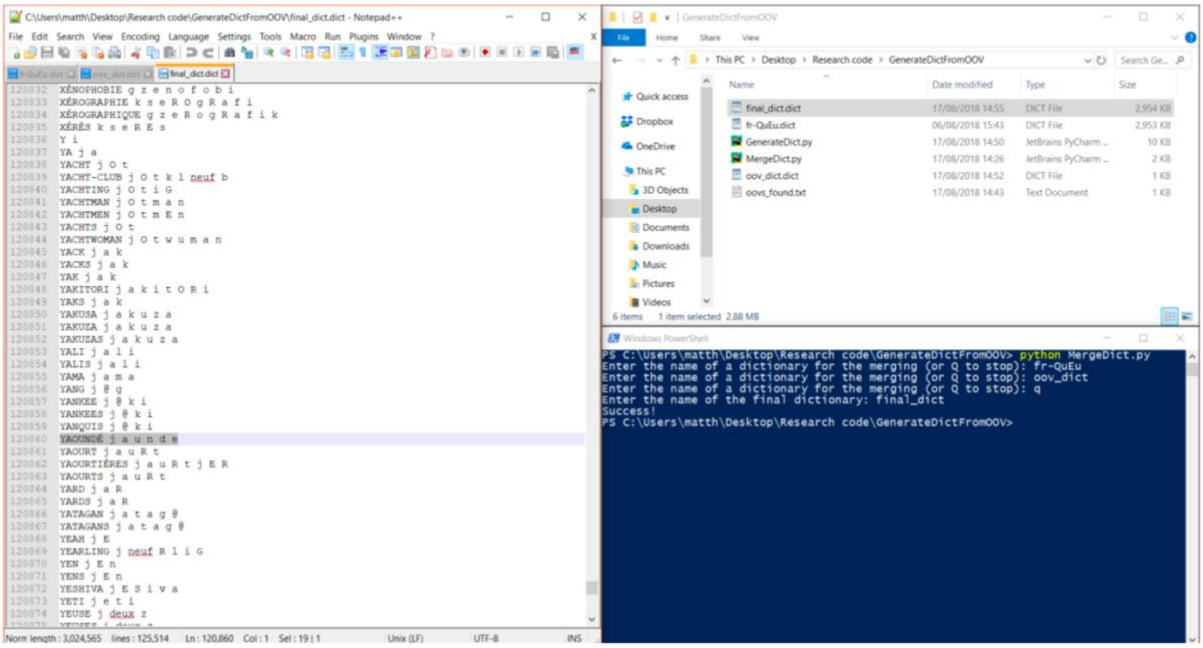
MergeDict.py combines multiple G2P dictionaries to get a broader dictionary with the OOVs included which can then be used by the MFA.

3.4.2 Requirements

- Python3 must be installed on your computer
- The G2P dictionaries to be merged must be in the same folder as MergeDict.py.

3.4.3 Running

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

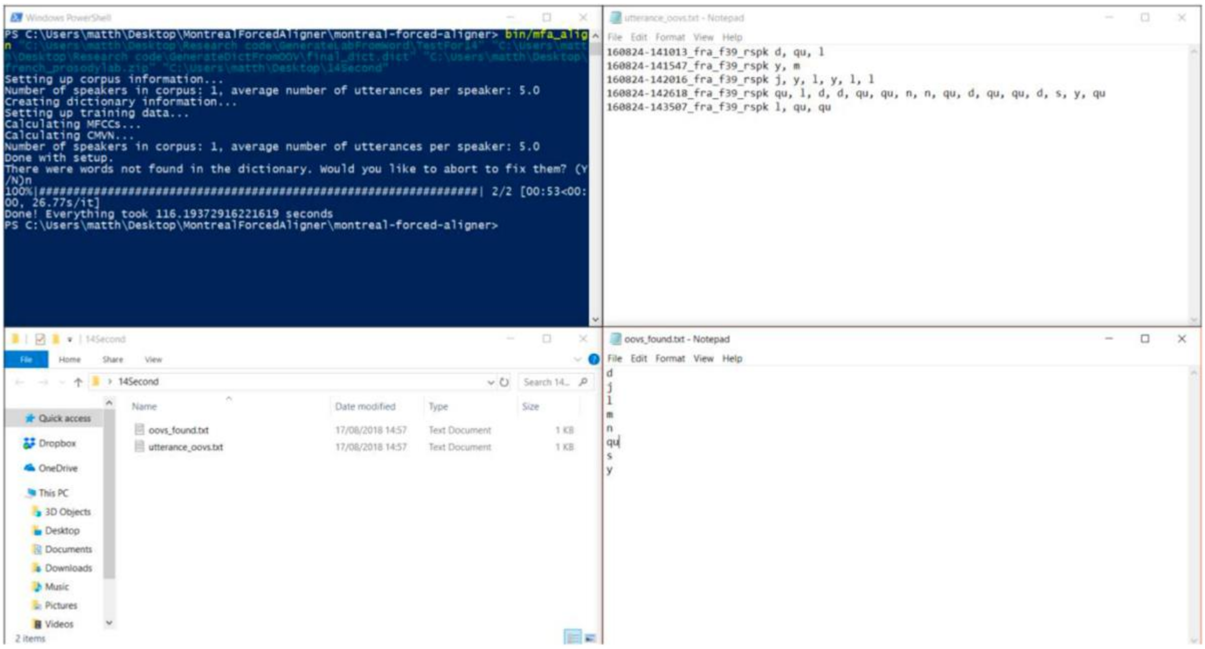

C:\Users\matth\Desktop\Research code\GenerateDictFromOOV\final_dict.dict - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
fr-QuEu.dict
oov_dict.dict
final_dict.dict
Norm length : 3,024,565 lines : 125,514 Ln : 120,860 Col : 1 Sel : 19 | 1 Unix (LF) UTF-8 INS
GenerateDictFromOOV
File Home Share View
This PC › Desktop › Research code › GenerateDictFromOOV
Search Ge...
Quick access
Dropbox
OneDrive
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
6 items 1 item selected 2.88 MB
Windows PowerShell

```
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
```

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

```
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>
```

You can now 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
```

4 Helper Functions

4.1 Introduction

The two functions below perform additional functions that helps clean the data.

4.2 MergeOOV.py

4.2.1 Introduction

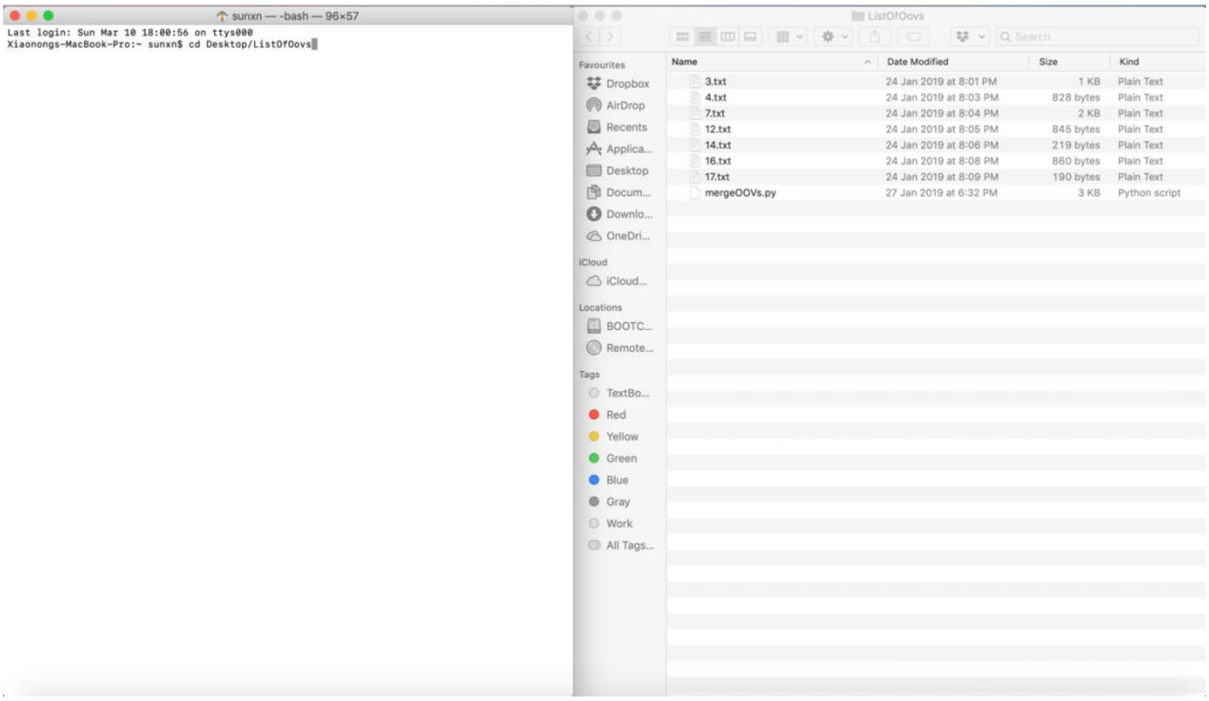
MergeOOV.py combines multiple OOV files into one OOV file without duplicates. This is helpful because the montreal forced aligner might be run a couple of times to account for different sampling rates.

4.2.2 Requirements

- Python3 must be installed on your computer
- The OOVs to be merged have to be in the same folder as MergeOOV.py

4.2.3 Running

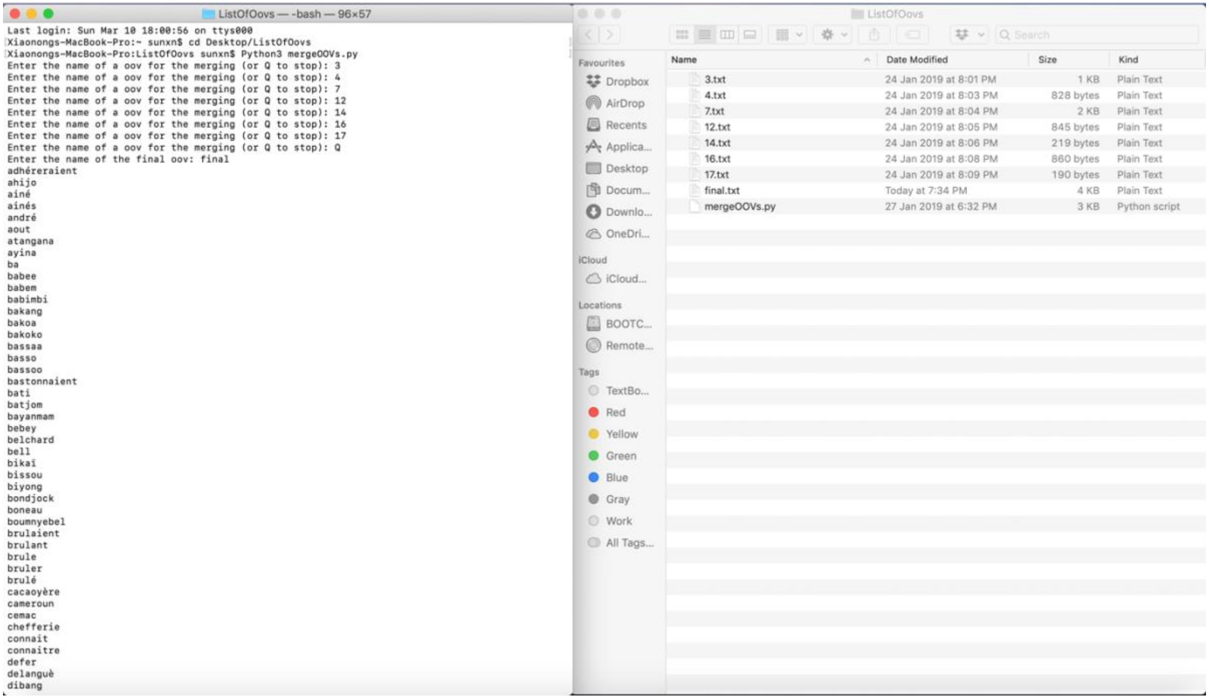
Have the required files ready and open the terminal. In this case, the files are the OOVs to be merged. Navigate to the folder using terminal.



```
Last login: Sun Mar 10 18:00:56 on ttys000
Xiaonongs-MacBook-Pro:~ sunxn$ cd Desktop/ListOfOovs
```

Name	Date Modified	Size	Kind
3.txt	24 Jan 2019 at 8:01 PM	1 KB	Plain Text
4.txt	24 Jan 2019 at 8:03 PM	828 bytes	Plain Text
7.txt	24 Jan 2019 at 8:04 PM	2 KB	Plain Text
12.txt	24 Jan 2019 at 8:05 PM	845 bytes	Plain Text
14.txt	24 Jan 2019 at 8:06 PM	219 bytes	Plain Text
16.txt	24 Jan 2019 at 8:08 PM	860 bytes	Plain Text
17.txt	24 Jan 2019 at 8:09 PM	190 bytes	Plain Text
mergeOOVs.py	27 Jan 2019 at 6:32 PM	3 KB	Python script

Run the MergeOov.py and follow the instructions.



```
Last login: Sun Mar 10 18:00:56 on ttys000
Xiaonongs-MacBook-Pro:~ sunxn$ cd Desktop/ListOfOovs
Xiaonongs-MacBook-Pro:ListOfOovs sunxn$ Python3 mergeOOVs.py
Enter the name of a oov for the merging (or Q to stop): 3
Enter the name of a oov for the merging (or Q to stop): 4
Enter the name of a oov for the merging (or Q to stop): 7
Enter the name of a oov for the merging (or Q to stop): 12
Enter the name of a oov for the merging (or Q to stop): 14
Enter the name of a oov for the merging (or Q to stop): 16
Enter the name of a oov for the merging (or Q to stop): 17
Enter the name of a oov for the merging (or Q to stop): Q
Enter the name of the final oov: final
adhéreraient
ahijo
ainé
ainés
andré
aout
atangana
ayina
ba
babee
babem
babimbi
bakang
bakoa
bakoko
bassaa
basso
bassoo
bastonnaient
bati
batjom
bayanmam
bebey
belchard
bell
bikaï
bissou
biyong
bondjock
boneau
boumnyebel
brulaient
brulant
brule
bruler
brulé
cacaoyère
cameroun
cemac
chefferie
connait
connaitre
defer
delanguè
dibang
```

Name	Date Modified	Size	Kind
3.txt	24 Jan 2019 at 8:01 PM	1 KB	Plain Text
4.txt	24 Jan 2019 at 8:03 PM	828 bytes	Plain Text
7.txt	24 Jan 2019 at 8:04 PM	2 KB	Plain Text
12.txt	24 Jan 2019 at 8:05 PM	845 bytes	Plain Text
14.txt	24 Jan 2019 at 8:06 PM	219 bytes	Plain Text
16.txt	24 Jan 2019 at 8:08 PM	860 bytes	Plain Text
17.txt	24 Jan 2019 at 8:09 PM	190 bytes	Plain Text
final.txt	Today at 7:34 PM	4 KB	Plain Text
mergeOOVs.py	27 Jan 2019 at 6:32 PM	3 KB	Python script

4.3 GenerateWav.py

4.3.1 Introduction

The purpose of GenerateWav.py is to create empty .wav files from a .xlsx file containing a digitized Basaa-French dictionary for alignment.

For each Basaa word within the .xlsx file, 4 .wav files are created. They are named by the following rule:

<Basaa word><Ending>.wav

<Ending> may be one of BW, FW, BE, or FE, respectively standing for Basaa Word, French Word, Basaa Example and French Example. Spaces are replaced by the underscore character _ .

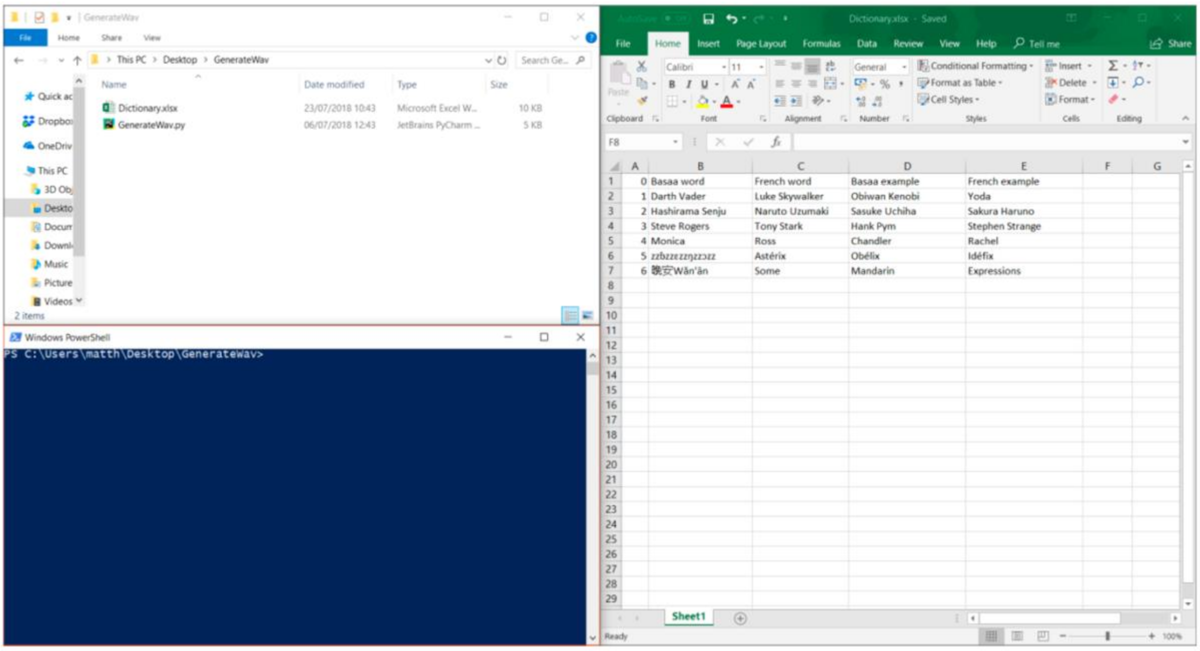
4.3.2 Requirements

- Python3 must be installed on your computer
- The Basaa-French dictionary is on the first sheet of the .xlsx file.
- The entries for the dictionary start on the second row, with the first row presumably containing the headings for each column.
- The .xlsx file must be in the same folder as GenerateWav.py
- The name given to new folder must be new
- The coloumn name must be alphabetic
- The elements in the “Basaa word” entry must not contain “\, /, :, *, ?, ", <, >, |”

- Certain words must be replaced with the following mapping
    - 'ɓ' to 'B'
    - 'ɛ' to 'E'
    - 'ŋ' to 'NG'
    - 'ɔ' to 'O'

4.3.3 Running

Have the required files ready and open your cmd/PowerShell/Terminal.

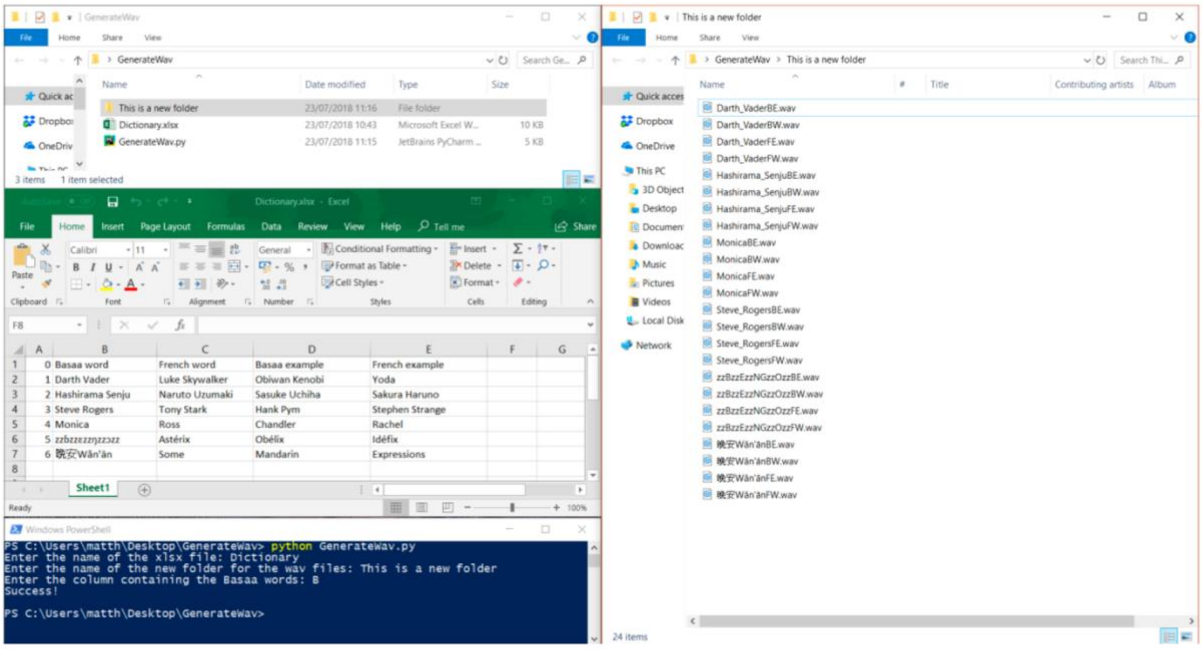


Name	Date modified	Type	Size
Dictionary.xlsx	23/07/2018 10:43	Microsoft Excel W...	10 KB
GenerateWav.py	06/07/2018 12:43	JetBrains PyCharm ...	5 KB

```
PS C:\Users\matth\Desktop\GenerateWav>
```

	A	B	C	D	E
1	0	Basaa word	French word	Basaa example	French example
2	1	Darth Vader	Luke Skywalker	Obiwan Kenobi	Yoda
3	2	Hashirama Senju	Naruto Uzumaki	Sasuke Uchiha	Sakura Haruno
4	3	Steve Rogers	Tony Stark	Hank Pym	Stephen Strange
5	4	Monica	Ross	Chandler	Rachel
6	5	zzɓzzɛzzŋzzɔzz	Astérix	Obélix	Idéfix
7	6	晚安Wǎn'ǎn	Some	Mandarin	Expressions

Run GenerateWav.py and follow the instructions. A new folder containing the .wav files should be created.



Name	Date modified	Type	Size
This is a new folder	23/07/2018 11:16	File folder	
Dictionary.xlsx	23/07/2018 10:43	Microsoft Excel W...	10 KB
GenerateWav.py	23/07/2018 11:15	JetBrains PyCharm ...	5 KB

```
PS C:\Users\matth\Desktop\GenerateWav> python GenerateWav.py
Enter the name of the xlsx file: Dictionary
Enter the name of the new folder for the wav files: This is a new folder
Enter the column containing the Basaa words: B
Success!

PS C:\Users\matth\Desktop\GenerateWav>
```

GenerateWav > This is a new folder

Darth_VaderBE.wav
Darth_VaderBW.wav
Darth_VaderFE.wav
Darth_VaderFW.wav
Hashirama_SenjuBE.wav
Hashirama_SenjuBW.wav
Hashirama_SenjuFE.wav
Hashirama_SenjuFW.wav
MonicaBE.wav
MonicaBW.wav
MonicaFE.wav
MonicaFW.wav
Steve_RogersBE.wav
Steve_RogersBW.wav
Steve_RogersFE.wav
Steve_RogersFW.wav
zzBzzEzzNGzzOzzBE.wav
zzBzzEzzNGzzOzzBW.wav
zzBzzEzzNGzzOzzFE.wav
zzBzzEzzNGzzOzzFW.wav
晚安Wǎn'ǎnBE.wav
晚安Wǎn'ǎnBW.wav
晚安Wǎn'ǎnFE.wav
晚安Wǎn'ǎnFW.wav

24 items

5 Glossary

G2P – Grapheme to Phoneme, see Additional Resources for definition of Grapheme and Phoneme
MFA – Montreal Forced Aligner
OOV – Out of Vocabulary
Pronunciation Dictionary – See section 2.1 for definition

6 Additional Resources

1. Acoustic Models - https://en.wikipedia.org/wiki/Acoustic_model
2. Graphemes, Phonemes definitions - https://prowritingaid.com/art/317/What-are-Phonemes%2C-Graphemes%2C-and-Digraphs.aspx
3. Montreal Forced Aligner - https://montreal-forced-aligner.readthedocs.io/en/latest/
4. PRAAT - http://www.fon.hum.uva.nl/praat/

7 Acknowledgements

The following script was written by Matthieu Chan Chee (matthieuchanchee@gmail.com), who was a Research Assistant under Dr Fatima Hamlaoui at the University of Toronto from June to August 2018:

MergeDict.py
GenerateWav.py

Along with the first version of

GenerateLab.py
GenerateDict.py

You can find the original version of these files in https://github.com/Dragionic/automate-alignment and https://github.com/Dragionic/wav-files-from-basaa-french-dict.