

Alignement de fichiers audios

1 Installation

Cette procédure n'est à faire qu'une fois, pour installer les logiciels nécessaires à l'alignement.

1.1 Installer numpy

Installer le module numpy avec la commande :
`pip3 install numpy`

1.2 Installer aeneas

Pour installer aeneas sur un ordinateur Linux :

1. Dans un terminal, taper :
`wget https://raw.githubusercontent.com/readbeyond/aeneas/master/install_dependencies.sh`
2. Quand la procédure est finie, taper :
`bash install_dependencies.sh`
3. Ouvrir un terminal dans le répertoire où on veut que le dossier de l'application soit mis (ex. : Documents)
4. Taper :
`git clone https://github.com/ReadBeyond/aeneas.git`
5. Taper :
`cd aeneas`
6. Taper :
`sudo pip3 install -r requirements.txt`
7. Taper :
`python3 setup.py build_ext --inplace`
8. Taper :
`python3 aeneas_check_setup.py`
9. Pour vérifier que le programme marche, taper :
`python3 -m aeneas.tools.execute_task`

Pour faire fonctionner le programme, il faudra ouvrir le terminal **dans le répertoire aeneas**.

1.3 Télécharger et installer Praat (optionnel)

Si vous voulez vérifier les alignements d'aeneas (fortement recommandé), vous devez installer Praat. Vous pouvez le télécharger ici : <https://www.fon.hum.uva.nl/praat>. Toutes les instructions pour l'installation sont dans les pages « Download ».

1.4 Récupérer les programmes

Récupérer les programmes du kit d'alignement automatique du Congrès :

- add_variety.py
- align_aeneas.py
- align_variety.py
- convert_json_textgrid.py
- rename_files.py
- split_audio.py

split_sentence.py

Copier coller le programme align_aeneas.py à l'intérieur du répertoire aeneas.

2 Préparer les fichiers

Préparer un répertoire avec tous les fichiers audios (format wav). Attention, il ne faut **pas de majuscules**, d'accents ou d'espaces dans les noms de fichiers ou de dossiers.

2.1 Renommer les fichiers si besoin

Si vous voulez renommer automatiquement tous les fichiers d'un dossier pour qu'il n'y ait plus de majuscules et caractères spéciaux, vous pouvez utiliser le programme rename_files.py avec la commande :

```
python3 rename_files.py repertoire_des_fichiers_a_renommer
```

Ce programme renommera chacun des fichiers du répertoire en enlevant les caractères spéciaux et en passant les majuscules en minuscules.

2.2 Formater les transcriptions

Pour chaque fichier audio, il faut un fichier de transcription en .csv. Il doit avoir le même nom (à part l'extension ; ex. : cercapais.wav et cercapais.csv). Dans le fichier, il doit y avoir une phrase par ligne, sans autre information. Attention, il ne faut **pas de majuscules**, d'accents ou d'espaces dans les noms de fichiers.

Si vous voulez automatiser le découpage par phrase à partir de fichiers **txt**, vous pouvez utiliser le programme split_sentence.py avec la commande :

```
python3 split_sentence.py repertoire_des_fichiers_a_decouper
```

Ce programme créera, pour chaque fichier txt du répertoire, un fichier csv avec une phrase par ligne.

3 Aligner

Copier le répertoire contenant les fichiers à l'intérieur du répertoire aeneas que vous avez créé pendant l'installation à l'étape 1.

Ouvrez un terminal dans le répertoire aeneas et lancez la commande :

```
python3 align_aeneas.py nom_du_repertoire_des_fichiers
```

Le programme crée des fichiers en .json qui se trouvent à présent dans le répertoire avec les fichiers audios et leurs transcriptions.

4 Convertir en TextGrid

Convertir les fichiers json en TextGrid avec la commande :
`python3 convert_json_textgrid.py repertoire_des_fichiers`

Il n'y a plus besoin que le répertoire des fichiers soit dans le répertoire aeneas.

On obtient des fichiers TextGrid qu'on peut ouvrir avec le logiciel Praat.

5 Correction des fichiers Praat

Cette étape est optionnelle mais fortement recommandée. Il faut surtout vérifier au début et à la fin du fichier, et quand il y a des blancs au milieu, car l'aligneur ne détecte pas les silences.

1. Ouvrir Praat.
2. Dans la fenêtre Praat Objects, cliquer sur « Open » puis « Read from file »
3. Naviguer jusqu'au répertoire avec les fichiers. Sélectionner le fichier wav et le fichier TextGrid. Cliquer sur « open ».
4. Dans la boîte sous « Objects : », sélectionner les deux fichiers
5. Cliquer sur « View & Edit »
6. Corriger : corriger les textes quand il n'y a pas écrit exactement ce qui est dit, déplacer les limites autour du texte quand elles ne sont pas au bon endroit, éventuellement ajouter un intervalle...
7. Sauvegarder le fichier TextGrid en écrasant l'ancien.

Si vous avez besoin d'un tutoriel pour utiliser Praat, vous pouvez aller à l'adresse :
<https://llacan.cnrs.fr/fichiers/manuels/Praat/PRAAT%20TutorialFR.pdf>.

6 Faire un fichier par phrase

Passer le programme `split_audio.py` avec la commande :
`python3 split_audio.py repertoire_des_fichiers`

On obtient dans un sous-dossier « exports » tous les fichiers sons correspondant chacun à une phrase. On a ensuite pour chaque fichier original un fichier csv mettant en correspondance le nom des fichiers sons avec la phrase qui y est prononcée.

7 Ajouter l'information de variété

7.1 Si toutes les phrases de tous les fichiers sont dans la même variété

Pour ajouter la variété à la fin de chaque ligne des fichiers de correspondance, utiliser le programme `add_variety.py` avec la commande :
`python3 add_variety.py repertoire_des_fichiers variete_des_phrases`

Attention, il faut indiquer le répertoire qui contient les fichiers découpés (le répertoire « **exports** » si vous ne l'avez pas renommé depuis l'étape 6).

On obtient, dans le dossier exports, des fichiers terminés par `_var.csv` qui contiennent :

- le nom du fichier audio
- la phrase transcrite
- la variété de la phrase

7.2 S'il y a plusieurs variétés dans les fichiers

Reprenez les fichiers csv créés à l'étape 2 et créez une 2^e colonne dans laquelle vous indiquerez, en face de chaque phrase, sa variété. Enregistrez ce fichier avec le même nom que le fichier original et **_var** à la fin (ex. : cercapais_var.csv), séparateur de champ 5 et pas de séparateur de chaîne de caractères. Laissez ce fichier dans le même répertoire que le fichier original (c'est à dire dans le répertoire parent du répertoire « exports »).

Passez le programme align_variety.py avec la commande :
`python3 align_variety.py repertoire_des_fichiers_exportés`

On obtient, dans le dossier exports, des fichiers terminés par _var.csv qui contiennent :

- le nom du fichier audio
- la phrase transcrite
- la variété de la phrase