Les Lundis de LIDILE – 5 novembre 2018

Initiation à ELAN pour l'annotation de corpus multimodaux

Leonardo CONTRERAS ROA

Table des matières

I – Créer un fichier de transcription / annotation

1. Ouvrez ELAN
2. **Fichier > Nouveau** → Sélectionnez le(s) fichier(s) média avec lesquels vous allez travailler et ajoutez-les à la liste de fichiers liés avec la touche 

    a. S'il s'agit de plusieurs vidéos (différentes prises d'une même scène) ou d'une vidéo et de sa piste audio dans un fichier à part, celles-ci doivent être synchronisées. Si ce n'est pas le cas, la synchronisation peut être effectuée sur Elan (cf. Modes, ci-dessous).
    b. S'il s'agit d'un Template (voir ci-dessous) il faut cocher l'option **Template** avant de le sélectionner).
3. Cliquez sur OK. Votre fichier est prêt pour commencer à être annoté.
4. Sauvegardez votre fichier - **Fichier > Enregistrer** (Ctrl+S)
5. Si jamais vous souhaitez changer ou ajouter un fichier vidéo ou audio, il est possible de revenir sur la fenêtre de fichiers liés en cliquant sur **Edition > Fichiers liés** (Ctrl+Alt+L)

II – Préparer la transcription / annotation (Template)

Avant de commencer à transcrire, il faut établir la structure du fichier de transcription/annotation. Cette structure peut être sauvegardée sous forme de TEMPLATE, et réutilisée pour plusieurs fichiers de transcription si l'on souhaite les transcrire de façon consistante.

1. Acteurs

ELAN organise la transcription en différents ACTEURS (ou *tiers*, en anglais). Chaque acteur est une nouvelle « ligne » dans la partition et peut représenter, entre autres :

a. Des locuteurs différents
b. Différents niveaux ou dimensions de transcription/annotation

Un acteur par défaut (*default*) est disponible au moment de la création d'un fichier de transcription.

2. Types

Les acteurs sont regroupés en TYPES. Les types servent à organiser les annotations à plusieurs acteurs et à plusieurs niveaux en établissant des règles ou des restrictions pour leur transcription. Avant de créer un acteur pour commencer la transcription, il faut créer les types selon lesquels ils seront organisés.

3. Stéréotypes

Les types, à leur tour, sont régis par des STEREOTYPES. Chaque stéréotype a des propriétés différentes :

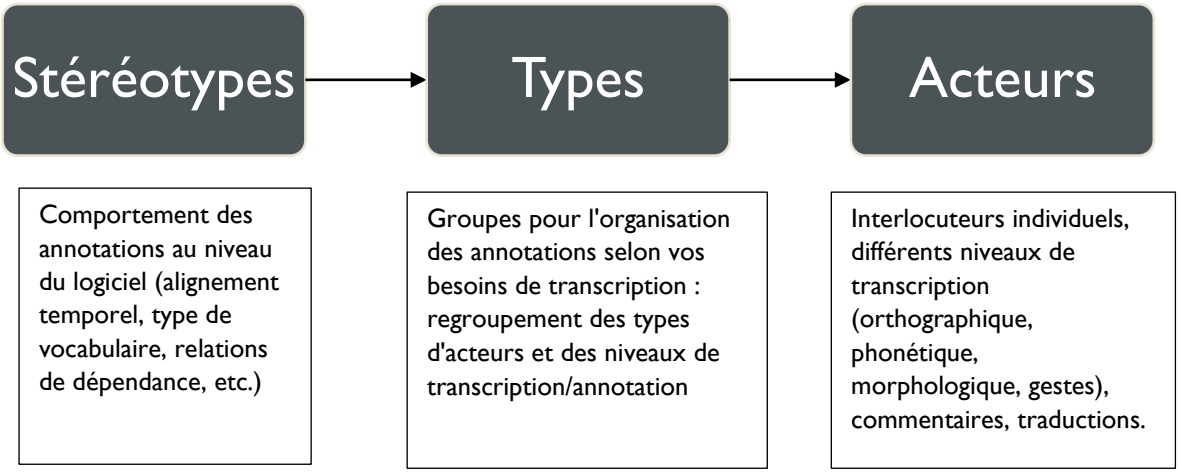
None → L'annotation est associée directement à l'axe du temps. Les acteurs appartenant à ce stéréotype sont indépendants. Exemple : Transcription orthographique à longs intervalles (divisée par des silences ou des pauses).

Time Subdivision → Permet de sous-diviser l'annotation d'un acteur parent en unités plus petites, obligatoirement contiguës, alignées sur le temps. Exemple : Une transcription orthographique sur peut être sous-divisée en mots individuels (ou *tokenisée*), ceux-ci alignés sur l'axe temporel à l'intérieur des limites de leur acteur parent.

Symbolic Subdivision → Similaire à Time Subdivision, mais les sous-unités ne sont pas alignées sur l'axe temporel et prennent toutes la même taille. Exemple : division des mots individuels en morphemes (qui ne sont pas liés à l'axe temporel).

Included In → Similaire à Time Subdivision, mais les éléments ne sont pas obligatoirement contigus, c'est-à-dire il peut y avoir des espaces vides entre eux.

Symbolic Association → Annotation sans sous-division possible d'un acteur parent. Exemple : Traduction du contenu orthographique d'un acteur parent dans une autre langue.



3.1. Hiérarchisation

Il existe des règles d'hiérarchisation :

- Les acteurs des stéréotypes « symboliques » (Symbolic Subdivision, Symbolic Association) ne peuvent pas être parents d'acteurs de type alignable (None, Time Subdivision, Included In).
- Une fois transcrits, les acteurs ne peuvent pas être changés à un autre type appartenant à un stéréotype différent.

L'hiérarchie des acteurs peut être visualisée en faisant **Clic droit** sur un acteur, puis sur **Affichage hiérarchisé**.

4. Créer un nouveau type

1. Pour créer un nouveau type, cliquez sur **Type > Ajouter nouveau type linguistique** (Ctrl+Maj+T)
2. Assignez un stéréotype au type (si vous ne savez pas encore combien de types utiliser, vous pouvez créer un type individuel pour chaque stéréotype)

4.1. Vocabulaire contrôlé (CV)

Certains acteurs auront un nombre limité d'entrées possibles (e.g. Annotation des gestes). Dans ce cas, pour accélérer l'annotation et réduire l'éventualité d'erreurs de transcription ou de coquilles, il est possible de créer des types à vocabulaire contrôlé.

1. Cliquez sur **Edition > Editer le vocabulaire contrôlé** (Ctrl+Maj+C)
2. Tapez le nom de la catégorie de vocabulaire contrôlé que vous souhaitez créer
3. Cliquez sur **Ajouter**
4. Vous pouvez ensuite ajouter des entrées individuelles pour cette catégorie.
5. Vous pourrez assigner cette catégorie de vocabulaire contrôlé lors de la création d'un nouveau type.

5. Créer un nouvel acteur

1. Pour créer un nouvel acteur, cliquez sur **Acteur > Ajouter nouvel acteur** (Ctrl + T)
2. Choisissez le nom de l'acteur, son parent (s'il y en a un) et son type. Les autres catégories sont facultatives.
3. Cliquez sur **Ajouter**

Vous pouvez modifier et supprimer les acteurs déjà existants en cliquant sur les onglets Modifier et Supprimer sur la fenêtre de création d'un acteur.

6. Exercice

Nous avons un fichier vidéo et la piste audio d'une conversation entre plusieurs personnes. Nous souhaitons transcrire et annoter les informations suivantes :

- Le contenu orthographique
- Les mots individuels
- La transcription phonétique des mots individuels
- Certaines informations paralinguistiques (rires, bâillements, toux, hésitations…)

Quels types et stéréotypes pour chacun de ces niveaux de transcription/annotation ?
Dans quelle hiérarchie les acteurs seront-ils organisés ?

7. Exporter votre Template

Une fois que tous les Types et acteurs ont été créés, vous pouvez exporter cette configuration pour la réutiliser lors de vos prochaines transcriptions. Pour ce faire :

1. Cliquez sur **Fichier > Enregistrer sous Template** (Ctrl+Alt+Maj+S)

Et voilà ! Vous pourrez sélectionner le *template* (fichier .etf) lors de la création d'un nouveau fichier, de la même façon où vous sélectionnez les fichiers audio et vidéo (cf. I – Créer un fichier de transcription / annotation)

III – Commencer à transcrire et à annoter

Une fois la structure interne de votre fichier de transcription établie, vous pouvez commencer à faire une première transcription orthographique. Il est possible de transcrire de deux façons différentes selon vos besoins, en utilisant des Modes différents.

1. Modes

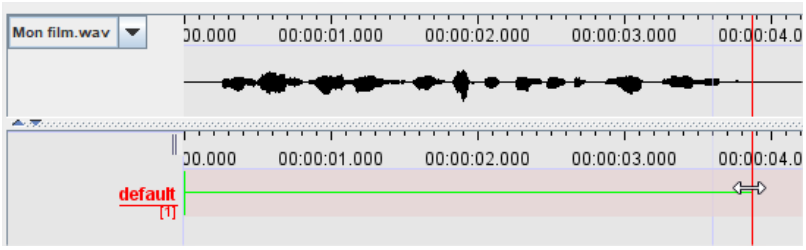
Pour changer de mode cliquez sur **Options**, où les différents modes sont affichés.

1.1. Mode Annotation

La première façon est à travers le mode ANNOTATION, qui est affiché par défaut. Il permet d'avoir une vue d'ensemble de tous les acteurs et de leur hiérarchisation. Les transcriptions/annotations déjà faites peuvent être affichées sous différentes formes en cliquant sur les onglets en haut à droite :

1. Pour ajouter une annotation sur un acteur alignable, sélectionnez une portion de la piste audio. Vous pouvez utiliser la forme d'onde du fichier son pour savoir où sont les silences.
2. Une fois la sélection faite, faites double-clic sur l'acteur où vous voulez ajouter une transcription. Cela créera un intervalle de transcription que vous pouvez remplir avec du texte :

3. Un intervalle déjà créé peut être déplacé en faisant **Alt+Clic** sur lui, et sa longueur peut être modifiée en faisant **Alt+Clic** sur l'un de ses bords :



4. Pour supprimer un intervalle, faites **Clic droit > Effacer Annotation**.

Cette façon de transcrire est la plus simple pour des fichiers courts et avec peu d'interlocuteurs. D'autres modes sont plus appropriés pour la transcription orthographique d'enregistrements plus longs.

1.2. Mode Segmentation

La deuxième façon de transcrire comporte deux étapes, dont la première est celle de SEGMENTATION. Dans ce mode, il est possible de créer les emplacements vides où seront insérées les transcriptions plus tard.

1. Contrairement au mode Annotation, le début et la fin de chaque intervalle se font individuellement en cliquant sur la ligne temporelle et en tapant **Entrée**, et non pas en sélectionnant un intervalle entier.
2. Pour effacer un intervalle, faites **Clic droit > Effacer Annotation** ou en cliquant sur lui et en tapant sur **Effacer** en même temps.

1.3. Mode Transcription

Une fois les intervalles créés, le mode transcription permet de faire une transcription rapide en écoutant le fichier son/vidéo et en vous servant du clavier pour passer rapidement d'un intervalle au prochain.

L'affichage est organisé en colonnes et chaque colonne correspond à un type d'acteur. Si vous avez créé un acteur différent pour chaque locuteur dans une conversation et qu'ils appartiennent tous au même type, vous les verrez tous simultanément sur la fenêtre transcription.

Chaque cellule correspond à un intervalle créé sur le mode Segmentation ou Annotation. Pour écouter le contenu d'une cellule, vous pouvez cliquer sur  ou taper sur la touche **Tabulation.** Pour passer d'une cellule à la prochaine, appuyez su **Alt + la Flèche** de direction souhaitée.

1.4. Mode Synchronisation

Le mode SYNCHRONISATION vous permet d'aligner des fichiers audio ou vidéo dont la lecture ne commence pas au même moment en établissant un décalage ou *offset* pour l'un d'entre eux.

2. Tokenisation

Un acteur qui contient une transcription orthographique et qui appartient à un stéréotype alignable sur le temps peut être sous-divisé automatiquement en *tokens* (mots individuels) sur un nouvel

acteur. Le processus ne fournit pas des tokens parfaitement alignés sur le temps et doivent être réajustés, mais ce processus est plus rapide que de faire une tokenisation manuelle.

1. Sélectionnez l'acteur à tokeniser.
2. Cliquez sur **Acteur > Tokeniser l'acteur**
3. Sélectioner les acteurs source et destination et le caractère de délimitation entre tokens (espace, par défaut)
4. Cliquez sur **Commencer**.

IV – Exportation et compatibilité

Une fois sauvegardé, votre fichier de transcription .eaf peut être exporté à d'autres formats de transcription, notamment .cha (format CHAT de CLAN) et .TextGrid (Praat). Inversement, Elan vous permet d'importer ce type de fichiers pour travailler sur un fichier audio déjà transcrit et l'enrichir ou le modifier. Pour exporter votre annotation vers d'autres formats,

Vous pouvez également exporter vos fichiers sous forme de tableaux au format .csv, compatible avec Excel et des éditeurs de texte.

1. Cliquez sur **Ficher > Exporter** vers et choisissez le format qui vous convient.