Préparation d'un fichier pour l'annotation dans Toolbox, à partir d'une transcription phrase par phrase d'un fichier audio dans Word

Transformation du fichier au format Toolbox

◊ Soit le fichier de transcription Word suivant (les phrases sont numérotées automatiquement : format, puces et numéro) correspondant à un enregistrement audio.

792. ʔèsásâ é să
793. lìkànɔ̀ɔ̂ kóà bóló wó tɛ̄ sèkò wó ʔɔ́ɔ́ lòtìó
794. mákàlà kɛ̀ɛ́, wá dɔ̄ nā kólòó ʔá gɔ́ɔ́ eee ! yélè ʔà nɔ́ɔ́ gbá
795. ʔá kɔ́tɔ̀ jɔ̀ɔ̀ ʔà ndá kɛ̀ wó ʔɔ́ɔ́ ná gɔ́ ɓó nɛ̀ɛ́ bóló ʔé tɛ́ sèkò ɸéé,
796. ɸéé kɔ̀ɔ́ tɔ́ ɸáà tɛ́ mò jó ɸé lɛ̀ pé sɔ́ bó mò kpîì békè náà ʔá tíndóà
797. ngɔ̀ ʔɔ́ɔ́ kɔ̀kɔ̀ɔ́ ʔá sóá ná kɛ̀ɛ́ kɔ́mbɛ̀ [aa] ʔá jó tɛ́ nɛ̀
798. bóló ngɛ́á ʔé ɸé sèkòó ʔá jɛ̀ ʔéɛ̀ kɔ̀ɔ́ ʔé mánàá ɸɛ́ nɛ̀
799. ɸé wó tɔ́ ɸó lɛ̀ jó kóà, lɛ̀ sɔ́ ʔɔ́ɔ́ bó, bóló ngɛ̀ɛ́, [e…] sèkò ngɛ́ɛ̀ ʔé ngɔ́ jáá békè náà,
800. ʔá kpílì ʔá tíndò wúú kpódè! bóló ngɛ́ sì, ndé ná… ná bá tɛ́ nà sɛ̀ wɔ́lɔ́
801. ʔá gɔ́á, ʔá gɔ́á jáá békè ná jɔ̀à ʔá tíndòá là bó
802. oo! yíɛ̀ mbìà wà yélè ʔá wɛ́ɛ̀ nɛ̀ ɸéé?

Il s'agit de remplacer chaque numéro de ligne par un marqueur de champ **\tx** précédé d'un marqueur de champ **\ref** vide

◊ Edition, Rechercher/remplacer tout, en cochant *Utiliser les caractères génériques*

Rechercher : `(<[0-9]@. {1;})(*^13)`

Remplacer : `^92ref ^13^92tx \2`

Résultat (après ajout manuel des 2 premières lignes)

```
\name BD-2011-C1
\id BD-2011-C1
\ref
\tx ʔèsásâ é să
\ref
\tx lìkànɔ̀ɔ̂ kóà bóló wó tɛ̄ sèkò wó ʔɔ́ɔ́ lòtìó
\ref
\tx mákàlà kɛ̀ɛ́, wá dɔ̄ nā kólòó ʔá gɔ́ɔ́ eee ! yélè ʔà nɔ́ɔ́ gbá
\ref
\tx ʔá kɔ́tɔ̀ jɔ̀ɔ̀ ʔà ndá kɛ̀ wó ʔɔ́ɔ́ ná gɔ́ ɓó nɛ̀ɛ́ bóló ʔé tɛ́ sèkò ɸéé,
\ref
\tx ɸéé kɔ̀ɔ́ tɔ́ ɸáà tɛ́ mò jó ɸé lɛ̀ pé sɔ́ bó mò kpîì békè náà ʔá tíndóà
…
```

◊ Enregistrer le fichier (Fichier, Enregistrer sous) au format *Texte Brut*, Encodage *Unicode/UTF-8* (par ex: BD-2011-C1.txt)

Ouverture du fichier transformé dans Toolbox

◊ Lancer **Toolbox**, puis Ouvrir le fichier préparé (Fichier, Ouvrir)

◊ Sélectionner le type *Text*

Le texte apparaît en entier dans une fiche Toolbox. Pour numéroter chaque phrase :

◊ Outils, **Découper/numéroter du texte**

◊ balise de référence : *ref* ; balise de texte : *tx*

◊ Utiliser le contenu du champ : *id* ou *name* puis faites OK

Les phrases sont maintenant numérotées.

Synchronisation des phrases avec un fichier Audio

Pour associer à chaque phrase le segment sonore du fichier audio correspondant au texte, nous allons dans un premier temps utiliser ELAN.

Segmentation du fichier audio dans ELAN

◊ Lancer **ELAN**

◊ Fichier, Nouveau

◊ Rechercher le fichier Audio correspondant au texte, le verser dans la fenêtre de droite avec le bouton >> ; OK

◊ Aller dans le menu **Options** : *Segmentation Mode*

◊ Sélectionner : *une frappe par annotation*

◊ Lancer la lecture et taper la touche *Entrée* à chaque fin de phrase.

◊ Une fois terminé, affiner le réglage des frontières entre phrases en plaçant la souris près d'une frontière, cliquez et déplacez. Attention de ne pas créer de trous (se placer dans le segment de gauche pour déplacer une frontière à droite, dans le segment de gauche pour déplacer une fronière à droite)

Contrôle de la segmentation

◊ Aller dans le menu **Options**, *Transcription Mode*

◊ Cliquer en face de 1 (*select a type*), Sélectionner *default-it*, Appliquer

Vous devriez avoir le même nombre de lignes dans ELAN que de références (\ref) dans le fichier Toolbox.

◊ Réduisez verticalement les fenêtres de Toolbox et de ELAN de façon à pouvoir les mettre l'une en dessous de l'autre

◊ Vérifier en cliquant sur chaque ligne de ELAN que le son correspond bien au contenu de la phrase de Toolbox de même numéro.

◊ Si nécessaire, retournez dans Options, *Segmentation Mode* pour ajouter ou fusionner des segments (clic-droit sur un segment : *split* pour dédoubler un segment, *merge* pour fusionner 2 segments)

◊ Enregistrer le fichier ELAN sous le même nom que le fichier Toolbox avec l'extension .eaf (BD-2011-C1.eaf)

Transfert des index temporels du fichier ELAN dans le fichier Toolbox

◊ Aller sur le site : *http://llacan.vjf.cnrs.fr/fichiers/Chanard/OutilsInfo/SynchroToolboxELAN.html*

◊ Rechercher votre fichier ELAN

◊ Rechercher votre fichier Toolbox

◊ Cocher la case *mot* si vous voulez ajouter une ligne d'annotation 'mot' (voir plus bas)

◊ Cocher la case *snd* si vous voulez pouvoir écouter le son dans Toolbox (outil, jouer un fichier Son)

◊ Lancer

◊ Afficher le fichier final puis enregistrez-le à partir du navigateur (Fichier, Enregistrer sous…) en lui donnant le même nom que le fichier Toolbox suivi de _EAF (ex: BD-2011-C1_EAF.txt)

Dans Toolbox, passer d'une fiche par texte à une fiche par phrase

Dans Toolbox, pour avoir une fiche par phrase au lieu d'une fiche par texte, il faut disposer d'un *type de base de données* ayant *\ref* comme marqueur d'enregistrement. Pour créer un tel type :

◊ Projet, Type de base de données, Sélectionner **Text** et faites *Copier*
Nom : **Texte** ; marqueur d'Enr : **ref**

◊ Fermer Toolbox

◊ Ouvrir le fichier Toolbox (BD-2011-C1_EAF.txt) avec Notepad++ (dans le navigateur Windows, clic-droit dessus, ouvrir avec… Notepad++)

◊ Remplacer *Text* par *Texte* dans la première ligne, enregistrer et fermer Notepad++

◊ Rouvrir le fichier (BD-2011-C1_EAF.txt) dans Toolbox

Maintenant chaque fiche contient une phrase numérotée, avec un marqueur ELANBegin et un marqueur ELANEnd contenant les index de début et de fin de la phrase du fichier Audio correspondant, ainsi qu'un marqueur ELANParticipant. Un champ \snd permet l'écoute dans Toolbox de chaque phrase (si le fichier audio n'a pas été déplacé de l'emplacement d'origine).

Avant de lancer l'interalignement dans Toolbox, il vous restera à vérifier que le lexique est bien associé au processus d'annotation.

Ce fichier Toolbox une fois annoté pourra être importé dans ELAN, chaque phrase (et ses annotations) sera couplée au segment sonore correspondant.

Ligne d'annotation \mot

Il peut être intéressant d'avoir d'une part une ligne de transcription \tx dont les mots ne seraient pas séparés par les espaces supplémentaires dus à l'interalignement, et d'autre part une ligne \mot qui, elle, sera la base de l'interalignement. Cette ligne *mot* peut être ajoutée à l'occasion du transfert des index par le formulaire précédent.

Si vous avez ajouté une ligne \mot vous devrez vous assurer que le *type* de base de données correspondant à votre fichier (type *Texte*) est bien configuré pour partir de la ligne *mot* comme base de l'interalignement. Pour le vérifier et éventuellement le modifier, faites

◊ Projet, Type de base de données

◊ Sélectionner le type *Texte*

◊ Modifier, Interalignement

La première ligne doit partir de 'mot'. Si elle part de 'tx', faites:

◊ Modifier, Sélectionner 'mot' dans la fenêtre déroulante *De la balise*

◊ Profitez-en pour vérifier que le bon lexique et le bon champ sont associés à cette ligne d'annotation puis faites OK, OK, Fermer (par la suite, vérifiez les autres lignes également).

Les annotations devraient dès lors être alignées correctement