

## 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. ʔèsásá é sǎ

793. likàndò kòà bóló wó tē sèkò wó ʔdò lòtió

794. mákàlà kèé, wá dō nā kólòó ʔá gòó eee ! yélé ʔà nòó gbá

795. ʔá kótò jòò ʔà ndá kè wó ʔdò ná gó bó nèé bóló ʔé té sèkò ʔéé,

796. ʔéé kòó tó ʔáà té mò jó ʔé lè pé só bó mò kpî békè náà ʔá tíndóà

797. ngò ʔdò kòkòó ʔá sóá ná kèé kómbè [aa] ʔá jó té nè

798. bóló ngéá ʔé ʔé sèkòó ʔá jè ʔéè kòó ʔé mánàá ʔé nè

799. ʔé wó tó ʔó lè jó kòà, lè só ʔdò bó, bóló ngèé, [e...] sèkò ngèé ʔé ngó jáá békè náà,

800. ʔá kpîli ʔá tíndò wúú kpódè! bóló ngé sî, ndé ná... ná bá té nà sè wóló

801. ʔá góá, ʔá góá jáá békè ná jòò ʔá tíndóà là bó

802. oo! yíè mbìà wà yélé ʔá wéè nè ʔéé?

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 ʔèsásá é sǎ

\ref

\tx likàndò kòà bóló wó tē sèkò wó ʔdò lòtió

\ref

\tx mákàlà kèé, wá dō nā kólòó ʔá gòó eee ! yélé ʔà nòó gbá

\ref

\tx ʔá kótò jòò ʔà ndá kè wó ʔdò ná gó bó nèé bóló ʔé té sèkò ʔéé,

\ref

\tx ʔéé kòó tó ʔáà té mò jó ʔé lè pé só bó mò kpî 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 frontière 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

Toolbox - [BD-2011-C14Ref\_EAF.txt]

Fichier Edition Base de Données Projet Outils Vérification Affichage Fenêtre Aide

[pas de filtre]

\ref	BD-2011-C14.034
. \ELANBegin	172.084
. \ELANEnd	173.914
. \ELANParticipa	SP
. \x	lää lè ʔá ndé ʔóó láá mò fé
. . \mot	lää lè ʔá ndé ʔóó láá mò fé
. . . \mb	lè -á lè ʔá ndé ʔóó lè -á mò fé
. . . . \ge	enfant -de IS voilà neg PASS enfant -de 2S disc
. . . . \ps	N -CON PRON PRES NEG MOD N -CON PRON DISC
. \snd	C:/Bertille-2011/RECORDER/contes/BD-2011-C14A.wav 172.084 173.914