

Charger l'ensemble des états d'un dossier dans Le Trameur à partir d'un fichier d'alignements

1. Récupérer le fichier d'alignement voulu au format XML :
 - soit en ligne sur le site du projet Écritures, onglet « Corpus du projet » :

Corpus du Projet

Le corpus Brouillons ADGB2008 (version électronique)

Le corpus de brouillons a été recueilli auprès de nos partenaires du S.A.F.E. de Caen. Nous les remercions pour leur participation et pour l'accès au corpus. Les textes sur lesquelles nous travaillons ont été anonymés. Compte tenu de la nature du corpus et de sa sensibilité sociale, celui-ci n'est pas en accès libre. Si vous êtes intéressés par les données ci-dessous, vous pouvez contacter georgeta.cislaru@univ-paris3.fr et serge.fleury@univ-paris3.fr.

Présentation du corpus

Disponible sur la page : http://syled.univ-paris3.fr/projet_innovant/ADGB2008/wip.html

Explorations du corpus (version 1)

Chaque dossier est constitué par les différents états de production du rapport visé.

Dossier n°1 : 24 états de fichier

Alignement des 24 états : V1, V2

- Chronologie de la variation sur les
- Chronologie des Segments répétés
- Chronologie de la variation des S

Dossier n°2 : 12 états de fichier

- Alignement des 12 fichiers états
- Chronologie de la variation sur les
- Chronologie des Segments répétés

Ce dossier est constitué par les 24 états

Ce dossier est constitué par les 12 états

Faire un clic droit sur le lien du fichier, puis sélectionner « Enregistrer la cible du lien sous » (ou dénomination similaire) pour sauvegarder le fichier localement.

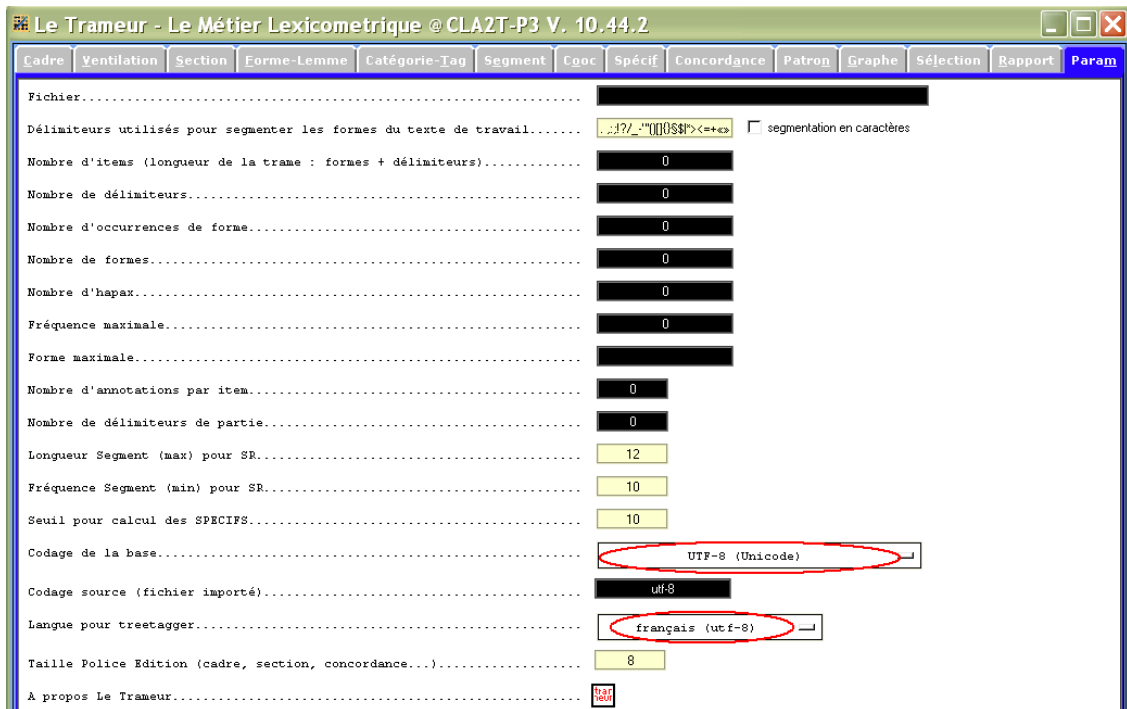
- soit directement dans le dossier partagé Dropbox « ANR-ECRITURES\CORPUS\Alignement » ;
2. Les fichiers d'alignements ont été produits par mkAlign et se trouvent au format TMX (*Translation Memory eXchange*), qui est une norme utilisée principalement dans les logiciels d'aide à la traduction. Il s'agit d'un document XML comme les autres et peut donc être chargé tel quel dans Le Trameur. Exemple de début de document (image tronquée à droite et en bas) :

```

<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet type="text/xsl" href="styles-tmx.xsl"?>
<tmx version="1.4">
<header adminlang="en" creationdate="20100329T171034Z" creationtool="mkAlign" creationont
<body>
<tu>
<tuv xml:lang="f1">
<seg>Prépa synthèse Anthony Viti. Plan Information concernant le placement. Connaissanc
</seg>
</tuv>
<tuv xml:lang="f2">
<seg>Plan Information concernant le placement. Connaissance de l'histoire familiale Rai
</seg>
</tuv>
<tuv xml:lang="f3">
<seg>Plan Information concernant le placement. Connaissance de l'histoire familiale Set
Après, un passage sur un groupe de MDEFC (groupe Mozaique), Anthony a été orienté vers
L'appel dénonce les relations complexes entre Anthony et sa mère.
Anthony subirait des mauvais traitements de la mère.
Suite au recueil, une démarche d'accompagnement de la famille a préconisé un éloignemer
</seg>
</tuv>

```

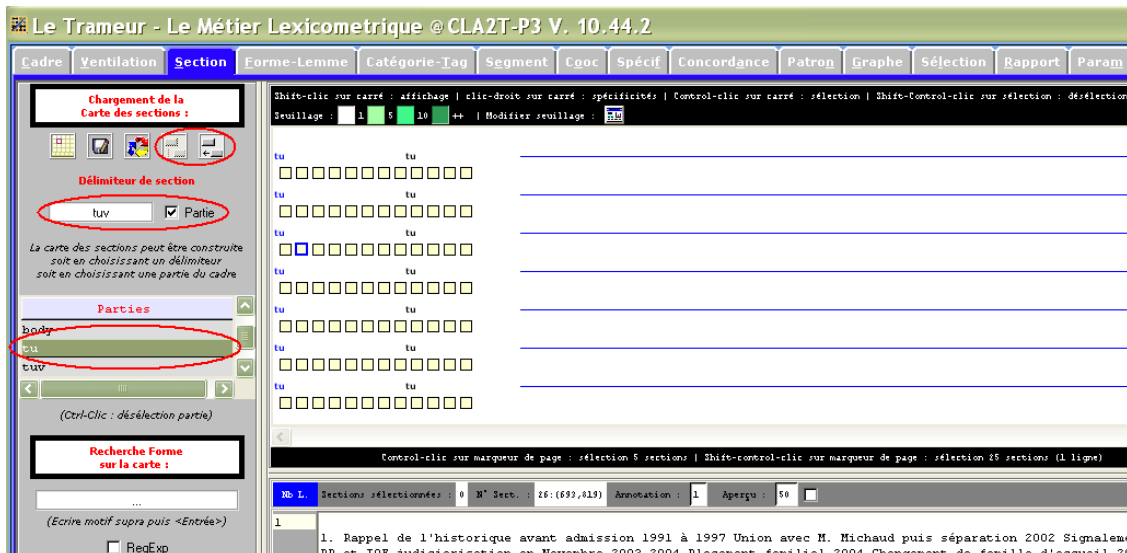
- le corps du texte se trouve à l'intérieur de la balise **<body>** ;
 - à l'intérieur de la balise **<body>** : les différents états d'un même « segment » (grain variable, typiquement de l'ordre de la phrase ou du paragraphe) se trouvent entre balises **<tu>** (pour *translation unit* — le format TMX est théoriquement prévu pour coder des traductions) ;
 - à l'intérieur d'une balise **<tu>** : les différents états d'un même segment se trouvent entre balises **<tuv>** (pour *translation unit variant*). Cette balise dispose d'un attribut **xml:lang** servant à définir la langue du segment. Dans notre cas il n'est pas question de langues, mais d'états, nommés ici f1, f2, f3, etc. ;
 - à l'intérieur d'une balise **<tuv>** : le texte du segment à proprement parler se trouve entre balises **seg**. Une balise **<tuv>** contient toujours une unique balise **<seg>**, et rien d'autre (dans les fichiers produits par mkAlign du moins). Utiliser les balises **<tuv>** ou **<seg>** dans Le Trameur reviendra donc au même dans notre cas.
3. Dans Le Trameur, onglet « Param », sélectionner **UTF-8** comme codage de la base (les fichiers TMX produits par mkAlign sont codés en UTF-8). Éventuellement, sélectionner utf-8 également dans la langue pour le TreeTagger, si l'étiquetage est nécessaire :



Puis charger la base comme à l'accoutumée par l'onglet « Cadre ».

4. Enfin, dans l'onglet « Section » :

- masquer les séparateurs de blocs de 5 sections et aligner les sections à gauche (faire un clic droit sur chacun des deux boutons en haut à gauche) ;
- entrer **tuv** comme délimiteur de section à la place du caractère « § », et cocher la case « Partie » ;
- sélectionner **tu** comme délimiteur de parties ;
- puis afficher la carte des sections :



La carte donne à présent une vue globale de l'ensemble des états du dossier. Chaque ligne correspond à un segment, et chaque colonne correspond à un état. Dans la capture ci-dessus par exemple, la section sélectionnée (le carré avec un contour bleu) correspond au deuxième état du troisième segment.