Tutoriel Prise en main d'Unitex-XAlign pour l'alignement de corpus

Denis Maurel
Université de Tours

La correction complète de ce tutoriel est disponible sur :

https://tln.lifat.univ-tours.fr/medias/fichier/correctionunitex-80jourschapitre2_1580220268618-zip?ID_FICHE=334600&INLINE=FALSE

Préparation

Installation

La première chose à faire est d'installer le logiciel libre Unitex. Il faut télécharger la version 3.2 ou une version ultérieure : https://unitexgramlab.org/. En cas de difficultés d'installation, le manuel est aussi accessible en ligne : https://unitexgramlab.org/releases/3.2/man/Unitex-GramLab-3.2-usermanual-fr.pdf. Il faut alors consulter les sections 1.3-1.7. Dans la phase d'installation, il faut choisir au moins les ressources françaises et anglaises.

Quelques mots rapides sur Unitex

Une fois le logiciel installé, le manuel Unitex est disponible via le menu *Help/Manuals*.

Ouvrir un texte : voir le manuel, section 2.4-2.5.5.

Créer un graphe : voir le manuel, section 5.2.

En bref, pour créer un graphe, il faut retenir les trois points suivants :

1. Créer (ou supprimer) un chemin entre deux boites : cliquer sur la première, puis sur la deuxième boite.
2. Créer une boite : clic-droit de la souris et choisir *Create box*. Si une boite est sélectionnée, un chemin sera automatiquement tracé entre la boite sélectionnée et la nouvelle boite.

3. Remplir une boite : écrire sur la barre de formule et valider. Si la barre de formule est totalement vide, la boite est supprimée. Le symbole *<E>* désigne l'élément vide (*empty*) et permet de créer une boite vide.

Préparation

Le plus simple est de télécharger le fichier :
https://tln.univ-tours.fr/medias/fichier/preparation-tutoriel-unitex-cassys-denis-maurel_1562936277186-zip?ID_FICHE=321996&INLINE=FALSE
et de le dézipper dans votre dossier personnel Unitex (les fichiers se placeront au bon endroit).

1 Création d'un fichier XML dans chaque langue et dans XAlign

Voir le manuel, section 10.1.

Se placer dans le dossier *French\Corpus\80JoursChapitre2* et ouvrir le fichier *80JoursChapitre2French.txt* avec un éditeur. Enregistrer ce fichier sous le nom *80JoursChapitre2French.xml*.

Ajouter des balises TEI au début et à la fin de ce fichier :

```
<tei>
<teiHeader/>
<body>
<text>
<div>Chapitre II…
…
une mécanique !"
</div>
</text>
</body>
</tei>
```

Se placer dans le dossier *English\Corpus\80JoursChapitre2* et ouvrir le fichier *80JoursChapitre2English.txt* avec un éditeur. Enregistrer ce fichier sous le nom *80JoursChapitre2English.xml*.

Ajouter les mêmes balises TEI au début et à la fin de ce fichier :

```
<tei>
<teiHeader/>
<body>
<text>
<div>Chapter II…
…
serving a machine."
</div>
</text>
</body>
</tei>
```

Se placer dans le dossier *XAlign\Corpus\80JoursChapitre2* et créer avec un éditeur le fichier ci-dessous, nommé *80JoursChapitre2FrenchEnglish.xml*.

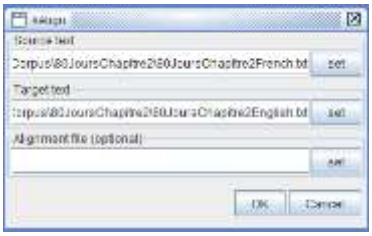
```
<tei>
<teiHeader/>
</tei>
```

2 Alignement automatique

Voir le manuel, section 10.2.

Ouvrons le menu *Xalign/Open Files* et remplissons les deux premières lignes avec :

1. Le fichier *French\Corpus\80JoursChapitre2\80JoursChapitre2French.txt*.
2. Le fichier *English\Corpus\80JoursChapitre2\80JoursChapitre2English.txt*.



Cliquons sur le bouton *OK*. Une fenêtre demande de sélectionner un fichier XML correspondant à la première ligne (*source file*). Cliquons sur *OK* et choisissons alors le premier fichier créé précédemment *French\Corpus\80JoursChapitre2\80JoursChapitre2French.xml*.

De même, une autre fenêtre demande de sélectionner un fichier XML correspondant à la deuxième ligne (*target file*). Choisissons alors le deuxième fichier créé précédemment *English\Corpus\80JoursChapitre2\80JoursChapitre2English.xml*.

Cliquons sur le bouton *OK*.

Sur la fenêtre suivante, qui présente les deux textes, cliquons sur le bouton *Align*. Une nouvelle fenêtre demande de sélectionner le fichier XML résultat. Choisissons alors le troisième et dernier fichier créé à l'étape précédente, à savoir le fichier *XAlign\Corpus\80JoursChapitre2\80JoursChapitre2FrenchEnglish.xml*.

Cliquons sur le bouton *Save alignment*.

Voir la vidéo : https://tln.lifat.univ-tours.fr/medias/video/xalign_1580219575777-mp4?ID_FICHE=334625&INLINE=FALSE.

Remarque :si on a fermé Unitex après avoir enregistré l'alignement, on peut l'ouvrir à nouveau en remplissant les trois lignes avec les fichiers XML :

3 Relecture et correction manuelle

Un premier alignement est fait. Il est à relire et à corriger :

- ajouter 4-6 ;

- supprimer 5-6 ; ajouter 5-7 ;
- supprimer 6-7 ; ajouter 6-8 ;
- supprimer 7-8 ; ajouter 7-9 ;
- supprimer 8-9 ; ajouter 8-10 ; ajouter 8-11 ;
- supprimer 9-10 ; ajouter 9-12 ;
- supprimer 10-11 ; ajouter 10-13 ;
- supprimer 11-12 ;
- supprimer 12-12 ; ajouter 12-14 ;
- supprimer 13-13 ; ajouter 13-14 ;
- etc.

Cliquons sur le bouton *Save alignment*.

En général, une phrase est traduite par une autre. Cependant :

- certaines phrases sont traduites par deux autres (2-2, 2-3) ;
- certaines ne sont pas traduites (11) ;
- plusieurs phrases peuvent être traduites par une seule (12-14, 13-14) ;
- des inversions sont possibles (30-30, 31-29).

4 Recherche effectuée sur un des deux textes

Recherchons sur le texte original (le français) le nom du serviteur *Passepartout*. Pour cela, cliquons sur le bouton *Locate…* (en bas à gauche de la fenêtre *XAlign*) et tapons *Passepartout* (au lieu d'un mot, on aurait pu choisir un graphe). Xalign crée un fichier qu'il place dans *French\Corpus\80JoursChapitre2*[1] et qu'il va renommer avec un suffixe : *80JoursChapitre2French_Xalign.txt*. Ce fichier va être traité par Unitex et le résultat va apparaître dans la fenêtre XAlign en sélectionnant *Matched sentences* en bas à gauche. On obtient alors juste les phrases correspondant à notre requête. Ce qui nous permet de constater qu'un nom propre peut être traduit par un pronom (20-21).

Il n'est pas possible de lancer une cascade, mais si le résultat est un texte balisé sans balises <s> ou <p>, il est possible de lancer XAlign sur ce texte. Voir à titre d'exemple, dans le dossier *French\Corpus\80JoursChapitre2*, le fichier *80JoursChapitre2FrenchBalise.txt* où les noms de personnes ont été balisés. Il est aussi possible pour ce fichier d'utiliser directement le fichier d'alignement réalisé précédemment en le plaçant directement sur la troisième ligne, puisque le balisage n'a pas modifié les phrases.

5 Entêtes TEI

Les fichiers XML soumis à XAlign ne contiennent pas d'entête[2], celles-ci sont donc à ajouter après l'opération d'alignement. Remarquons que si nous en plaçons une, elle sera effacée. C'est donc inutile.

[1] Si on avait lancé la recherche à droite, le fichier aurait évidemment été créé dans le dossier *English\Corpus\IdcpAlign*.
[2] Excepté la balise *<teiHeader/>*.

Cette entête peut donc être ajoutée ensuite. À titre d'exemple, les entêtes des fichiers *French\Corpus\80JoursChapitre2\TEeiHeaderChapitre2French.xml*
English\Corpus\80JoursChapitre2\TEeiHeaderChapitre2English.xml

peuvent être copiées-collées dans les fichiers résultats.