Ressources linguistiques
M1 d'informatique

TP 1. Prise en main d'Unitex

Cette séance est consacrée à la prise en main du système Unitex. Nous verrons les différentes étapes de prétraitement d'un texte. Nous rechercherons ensuite des motifs dans le texte à l'aide d'expressions régulières puis à l'aide de graphes.

Exercice 1. Prétraitement du texte

1.1. Téléchargez le manuel d'Unitex en français ou en anglais :
https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-fr.pdf

Lancez Unitex depuis votre session Linux en choisissant Unitex dans le menu Education [1]. La première fois que vous lancez Unitex, il affiche un message qui vous indique quel répertoire vous servira de répertoire Unitex personnel. Il ne devra servir qu'à ça, car chaque sous-répertoire qu'il contiendra doit correspondre à une langue.
Les 3 principaux menus d'Unitex sont :
- Text pour les textes, qui doivent être au format Unicode ;
- DELA pour les dictionnaires au format DELA (Dictionnaires Électroniques du LADL) ;
- FSGraph pour les graphes au format .grf (le format .fst2 est celui des graphes compilés).

L'ouverture et le prétraitement du texte sont détaillés dans les sections 2.4 et 2.5 du manuel.
Ouvrez le texte 80jours.txt :
- dans le menu Text, choisissez Open ;
- dans le menu Files of Type, sélectionnez Text Files ;
- sélectionnez ensuite le fichier 80jours.txt ;
- cliquez sur Open.

Une fois que vous avez ouvert un texte, Unitex vous propose de lui appliquer une suite de prétraitements. À la question Do you want to preprocess the text ?, répondez Yes. Unitex ouvre alors une fenêtre intitulée Preprocessing & Lexical parsing dans laquelle vous pouvez paramétrer les prétraitements que vous souhaitez appliquer au texte. Dans la section Preprocessing, assurez-vous que les choix suivants sont cochés :

Apply graph in MERGE mode

en vérifiant que c'est bien le graphe Graphs/Preprocessing/Sentence/Sentence.grf qui est sélectionné : ce graphe, appliqué en mode Merge, est utilisé pour découper le texte en phrases. Il insère un marqueur {S} au début de chaque phrase du texte.

Apply graph in REPLACE mode

en vérifiant que c'est bien le graphe Graphs/Preprocessing/Replace/Replace.grf qui est sélectionné : ce graphe, appliqué en mode Replace, est utilisé pour remplacer des formes dans le texte, en particulier pour normaliser les formes non ambigües telles que les élisions (par exemple, la négation *n'* est remplacée par *ne*) et les contractions (par exemple, *auquel* est remplacé par *à lequel*).

[1] En cas d'échec, placez-vous dans le répertoire /usr/local/apps/Unitex/App/, puis lancez avec Java le .jar installé à l'adresse /usr/local/apps/Unitex/App/Unitex.jar.

Dans la section Lexical Parsing, cochez le choix suivant :

Apply All default Dictionaries

L'application des dictionnaires DELA à votre texte vous permet de le découper en unités lexicales et de créer des dictionnaires du texte.
Cliquez sur GO pour lancer le prétraitement. Le résultat du prétraitement est un nouveau fichier 80jours.snt et un répertoire 80jours_snt qui se trouvent dans le même répertoire que le fichier 80jours.txt. Le répertoire 80jours_snt contient entre autres les dictionnaires du texte : dlf pour les formes simples, dlc pour les formes composées et err pour les mots non trouvés dans les dictionnaires utilisés. Dans la fenêtre Unitex, vous pouvez visualiser la liste des unités lexicales (tokens) trouvées dans le texte ainsi que la liste des mots simples, composés et inconnus trouvés (Word Lists). Vous pouvez aussi ouvrir les fichiers dlf, dlc et err avec Open dans le menu DELA.

Ouvrez, avec Open dans le menu FSGraph, les graphes Sentence.grf et Replace.grf que vous avez appliqués au texte. Parcourez-les, ainsi que leurs sous-graphes, tout en observant dans le texte les effets de leur application. (Attention, FSGraph est un éditeur de graphes : si vous cliquez sur les nœuds, vous risquez de modifier les graphes.) Pour ouvrir un sous-graphe, cliquez sur un nœud grisé avec le bouton du milieu. Les **sorties** sont ce qui apparait au-dessous des nœuds. Ce sont des mots, des codes ou des commandes. Il existe 3 modes d'utilisation des sorties :
- Merge (« fusionner ») permet d'insérer les séquences produites par les sorties ;
- Replace permet de remplacer les séquences reconnues par les séquences produites ;
- le troisième mode ignore les sorties.
Notez ce que vous observez.

1.2. Dans le menu Text, cliquez sur Construct FST-Text... pour construire l'automate du texte (on peut aussi lancer cette commande au moment du prétraitement : pour cela, lancez le prétraitement avec Preprocess Text dans le menu Text, et cochez Construct Text Automaton). Ouvrez ensuite l'automate du texte et parcourez-le phrase par phrase. L'automate du texte est-il toujours acyclique ? La construction de l'automate du texte est détaillée dans le chapitre 7 du manuel.

Notez ce que vous observez.

Si vous voulez rouvrir un texte sur lequel le prétraitement a déjà été fait, ce n'est pas la peine de le refaire. Unitex réutilisera le résultat du prétraitement. Pour rouvrir le texte de cette façon,
- dans le menu Text, choisissez Open ;
- dans le menu Files of Type, sélectionnez Unitex Text Files ;
- sélectionnez ensuite le fichier 80jours.snt.

Exercice 2. Recherche de motifs : expressions régulières

2.1. Pour chercher dans le texte un motif défini par une expression régulière, allez dans le menu Text et choisissez Locate Pattern. Dans la fenêtre qui s'ouvre, cochez Index all utterances in text, sinon les résultats de vos recherches de motifs seront limités aux 200 premières occurrences.

Parcourez les sections 4.2 et 4.3 du manuel pour vous familiariser avec la fonctionnalité Locate Pattern d'Unitex, et testez les exemples d'expressions régulières qui sont fournis : les tokens,

les masques lexicaux, etc. Les tableaux de la section 3.1.3 présentent une liste non exhaustive des codes de catégories que vous pourrez utiliser dans vos expressions régulières, respectivement : les codes grammaticaux (tableau 3.1), les codes sémantiques (tableau 3.2) et les codes flexionnels (tableau 3.3). Ces codes ne fonctionnent que si les dictionnaires ont été appliqués au texte. Si vous avez suivi les instructions lors du prétraitement, ils l'ont été [2]. Les résultats de vos requêtes doivent s'afficher avec un contexte d'une ligne chacun.

Notez vos requêtes, les résultats, et ce que vous observez.

2.2. Vous allez créer vos propres expressions régulières pour rechercher des motifs dans le texte. Notez les éventuelles incohérences que vous pourrez remarquer dans les résultats de vos requêtes et expliquez-en les causes.

Recherchez les motifs suivants dans le texte :

(a) tous les adjectifs
(b) toutes les occurrences de la forme *pouvoir*
(c) toutes les occurrences de la forme *savons*
(d) toutes les formes fléchies dont la forme canonique est le verbe *pouvoir*

Notez vos requêtes, les résultats, et ce que vous observez.

Exercice 3. Recherche de motifs : grammaires locales

Le but de l'exercice est de construire une grammaire de détection des expressions d'horaire dans un texte.
Pour ouvrir un nouveau graphe, on utilise New dans le menu FSGraph. L'édition de graphes est détaillée dans le chapitre 5.2 du manuel. Voici la signification des principaux symboles :
- le <E> représente le mot vide ε ;
- le + sépare les différentes lignes à l'intérieur d'un nœud ;
- le : sert à introduire un appel à un sous-graphe ;
- le / indique le début de la sortie dans un nœud.

Vous pourrez tester vos graphes avec Locate Pattern (sélectionnez le graphe à appliquer au texte en cliquant sur Set). Vous avez intérêt à sauvegarder vos graphes dans le répertoire Graphs : c'est là qu'Unitex ira les chercher par défaut.

3.1. Recherchez tous les nombres dans le texte à l'aide d'une expression régulière dans Locate Pattern. Notez votre requête, les résultats et ce que vous observez.

3.2. Utilisez la concordance obtenue pour construire un graphe horaire.grf qui reconnait toutes les expressions d'horaire du texte. (Il y a beaucoup de façon d'écrire un horaire, mais vous pouvez vous limiter aux styles utilisés dans le *Tour du monde en 80 jours*.) Copiez votre graphe, notez les résultats et ce que vous observez.

[2] Sinon, vous pouvez vous rattraper en utilisant le menu Text > Apply Lexical Resources.