

Travaux Pratiques de linguistique

Ressources linguistiques

—Master 1—

Unitex

Cette séance est consacrée à la prise en main du logiciel Unitex. Nous verrons les différentes étapes de traitement des textes.

► Exercice 1. Logiciel Unitex

1. Télécharger la version 1.2 d'Unitex dans votre répertoire principal :
<http://www-igm.univ-mlv.fr/~unitex>.
Penser à télécharger également le manuel d'Unitex (pour la version en français :
<http://www-igm.univ-mlv.fr/~tolone/Linguistique/ManuelUnitex.pdf>).
2. Installer Unitex avec les commandes suivantes :
 - `unzip Unitex_1.2.zip -d Unitex`
 - `cd Unitex/Src/C++`
 - `make install`
3. Lancer Unitex sur la langue française :
 - `cd ../../App`
 - `java -jar Unitex.jar &`

Vous avez les 3 menus suivants :

- "Text" pour les textes au format Unicode ;
- "DELA" pour les dictionnaires au format DELA (Dictionnaires Électroniques du LADL) ;
- "FSGraph" pour les graphes au format .grf et .fst2 pour les graphes compilés.

► Exercice 2. Construction de graphes

Ouvrir un nouveau graphe avec "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 de mot vide epsilon ;
- le + sépare les différentes lignes des boîtes ;
- le : sert à introduire un appel à un sous-graphe ;
- le / indique le début de la sortie dans une boîte.

Il existe 3 modes d'utilisation des sorties :

- **MERGE** : 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.

1. Écrire une grammaire de normalisation permettant de remplacer toutes les occurrences de *l'on* par *on*. Enregistrer le graphe dans `unitex/French/Graphs/Preprocessing/Replace/` et le compiler.

► Exercice 3. Chargement d'un texte

Voici les différentes étapes du chargement du texte *Le tour du monde en 80 jours* :

- prétraitement du texte : normalisation et découpage en phrases ;
- découpage en unités lexicales ;
- application de dictionnaires ;
- construction de l'automate du texte.

1. Ouvrir les 2 graphes suivants avec "Open..." dans le menu "FSGraph" et explorer leurs sous-graphes :
 - `unitex/French/Graphs/Preprocessing/Replace/Replace.grf` qui permet de normaliser le texte ;
 - `unitex/French/Graphs/Preprocessing/Sentence/Sentence.grf` qui permet de découper le texte en phrases.

2. Ouvrir le texte `unitex/French/Corpus/80jours.txt` (de type Raw Unicode Texts) avec "Open..." dans le menu "Text".

Lancer le prétraitement du texte en appliquant `Sentence.grf`, la grammaire faite à l'exercice précédent à la place de `Replace.grf` (de type Unicode Graphs alors que les `.fst2` sont de type Unicode Compiled Graphs), ainsi que les dictionnaires par défaut. Vous obtenez la liste des unités lexicales (`tokens`) trouvées dans le texte ainsi que la liste des mots simples, composés et inconnus trouvés.

Le répertoire `unitex/French/Corpus/80jours_snt` contient toutes ces données.

Vous pouvez ouvrir les fichiers `dlf`, `dlc` et `err` avec "Open" dans le menu "DELA".

L'ouverture et le prétraitement d'un texte sont détaillés dans le chapitre 2.4 et 2.5 dans le manuel.

3. Construire l'automate du texte ("Construct FST-Text..." dans le menu "Text" ou relancer le prétraitement avec "Preprocessing Text..." dans le menu "Text" en construisant l'automate du texte). L'automate du texte est **acyclique**.

La construction de l'automate du texte est détaillée dans le chapitre 7.1 et 7.2 du manuel.

► Exercice 4. Recherche de motifs élémentaires

1. Rechercher le motif `parler` en cliquant sur "Locate Pattern..." dans le menu "Text".
2. Même question avec le motif `<parler>`.
3. Même question avec le motif `<V:P3p>`.
4. À quoi correspondent les motifs précédents ?