UNITEX 3.2

Abdelhalim hafedh Dahou

Introduction

Unitex est un ***ensemble de logiciels*** permettant de ***traiter des textes*** en langues naturelles en utilisant ***des ressources linguistiques***.

- Dictionnaires électroniques.
- Grammaires.
- Tables de lexique-grammaire.

Il faut noter que Unitex serait ***inutile*** sans les précieuses ***ressources linguistiques*** qu'il renferme.

Introduction

Unitex est un logiciel ***libre*** et ***multi-plateformes*** capable de fonctionner aussi bien sous Windows que sous Linux ou OS X. Le code source des programmes et les ressources linguistiques sont distribué avec le logiciel sous la licence LGPL et LGPLLR.

Afin de pouvoir utiliser unitex, il faut préalablement installer un environnement d'exécution JRE (Java Runtime Environment) et nécessairement une version 1.7 (ou plus récente) du java sinon Unitex se bloquera après que vous ayez choisi votre langue de travail.

Site officiel du unitex est le suivant : http://releases.unitexgramlab.org/latest-stable

Installation sous Linux

Pour installer le logiciel sous l'environnement linux , il faut appliquer les étapes suivant :

1. ***Télécharger*** le fichier suivant : Unitex-GramLab-3.2-linux-x86_64.run
2. Donnez-lui les ***droits d'exécution***, par exemple:

   chmod a+x Unitex-GramLab-3.2-linux-i686.run
3. L'exécution du fichier ***.run*** :

   /Unitex-GramLab-3.2-linux-i686.run
4. Dans le répertoire ***$HOME***, vous trouverez un répertoire personnel de travail appelé ***/unitex***,

Chargement d’un texte

Parmi les fonctionnalités d’Unitex est la recherche d’expressions ou bien des motifs dans des textes. Pour cela, il faut passer par les étapes suivants avant d’effectuer des recherches sur les textes :

1. Sélection de la langue.
2. Format des textes.
3. Édition de textes.
4. Ouverture d’un texte.
5. Prétraitement du texte.

Dictionnaires

Les dictionnaires électroniques utilisés par Unitex utilisent le formalisme DELA (Dictionnaires Electroniques du LADL). Ce formalisme permet de décrire les entrées lexicales simples et composées d'une langue en leur associant de façon optionnelle des informations grammaticales, sémantiques et flexionnelles.

1. DELAF.
2. DELAS.

Nous utiliserons les termes DELAF et DELAS pour désigner les deux sortes de dictionnaires que leurs entrées soient simples, composées ou mixtes.

Dictionnaires (DELAF)

Syntaxe d'une entrée

Une entrée d'un DELAF est une ligne de texte terminée par un retour à la ligne qui respecte le schéma suivant :

mercantiles,mercantile.A+z1:mp:fp/ceci un exemple

3\,1415,PI.NOMBRE

grand=mères,grand=mère.N:fp

Nous utiliserons les termes DELAF et DELAS pour désigner les deux sortes de dictionnaires que leurs entrées soient simples, composées ou mixtes.

Dictionnaires (DELAS)

Le format des DELAS est très similaire à celui des DELAF. La différence est qu'on ne mentionne qu'une forme canonique suivie de codes grammaticaux et/ou sémantiques. La forme canonique est séparée des différents codes par une virgule.

Voici un exemple d'entrée :

cheval,N4+Anl

Dictionnaires : Priorités

Pour éliminer certaines ambiguïtés lors de l'application des dictionnaires , on utilise les règles de priorité suivant:

- Les dictionnaires dont les noms terminent par - ont la priorité la plus **grande**.
- Ceux dont le nom se termine par + ont la priorité la plus **faible.**
- Les autres dictionnaires sont appliqués avec une priorité moyenne.

Dico ex.snt alph.txt ctr+.bin cities-.bin rivers.bin regions-.bin

Dictionnaires : Règles d’application

Outre la règle de priorités, l’application des dictionnaires s’effectue en respectant les

majuscules et les espaces.

- s’il y a une majuscule dans le dictionnaire, alors il doit y avoir une majuscule dans le texte.
- s’il y a une minuscule dans le dictionnaire, il peut y avoir soit une minuscule soit une majuscule dans le texte.
- pour qu’une séquence du texte soit reconnue par une entrée de dictionnaire, elle doit avoir exactement les mêmes espaces

Recherche d’expressions rationnelles

Unitex nous donne la possibilité d’utiliser les *expressions rationnelles* dans leur environnement pour rechercher des motifs simples. Une *expression rationnelle*, ou *expression régulière*, peut-être :

- une unité lexicale (*livre*) ou un masque lexical (*<manger.V>*)
- la concaténation de deux expressions rationnelles (je mange)
- l’union de deux expressions rationnelles (Pierre+Paul)
- l’étoile de Kleene d’une expression rationnelle (très*)
- une position particulière du texte : le début {^} ou la fin {$}

Les grammaires locales

Les grammaires locales sont un moyen puissant de représenter la plupart des phénomènes linguistiques.

Les grammaires Unitex sont des variantes des grammaires algébriques, également appelées grammaires hors-contexte. Une grammaire algébrique est constituée de règles de réécriture.

$$S \rightarrow aS$$

$$S \rightarrow \varepsilon$$

on appelle l'ensemble des mots reconnus par ces grammaires *langage d'une grammaire*.

Contextes

Les graphes d’Unitex sont des grammaires algébriques. Elles sont également appelées grammaires hors-contexte, car lorsque l’on souhaite reconnaître une séquence A, on ne tient pas compte du contexte dans lequel A apparaît. Par exemple, il est impossible de rechercher avec un graphe normal toutes les occurrences du mot président, sauf celles qui sont suivies par of the republic.

- Contextes droits.
- Contextes gauches.