Linguistique de corpus 2

Outils et méthodes de traitement de corpus 2

Patrick Paroubek

LIMSI-CNRS
Dépt. CHM - Groupe LIR
Bât. 508 Université Paris XI, 91403 Orsay Cedex
pap@limsi.fr

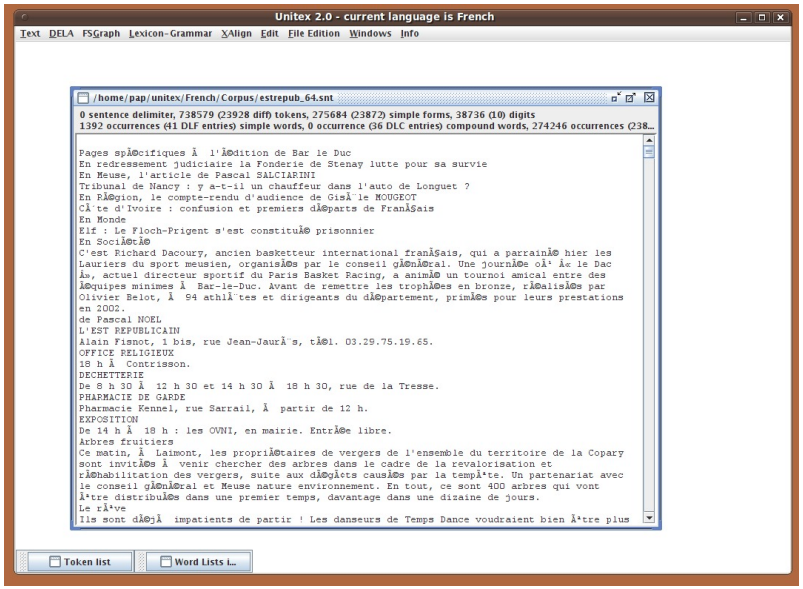
mercredi 01 février 2017 / L3 - Cours S2-1

Introduction à Unitex

La boîte à outils Unitex permet d'effectuer des analyses linguistiques de documents avec une interface graphique de haut-niveau.
Unitex et NOOJ sont deux plateformes issues des mêmes travaux (Maurice Gross) et comparables en fonctionalités.
L'importation d'un texte dans Unitex propose par défaut une segmentation en mot et en phrase ainsi qu'un ensemble de pré-traitements de normalisation (par ex. encodage des caractères).

Introduction à Unitex

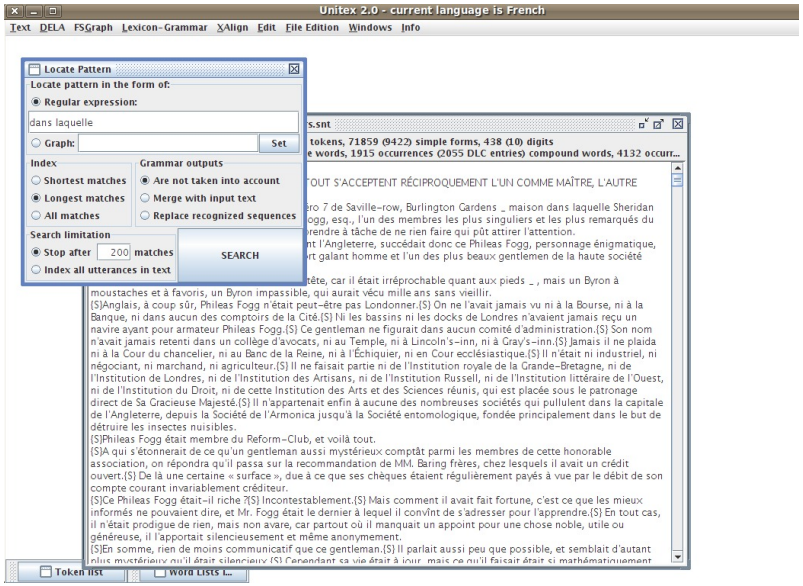
Charger un texte dans la boite à outils :



En activant l'item de menu : Text/Open... Le chargement produit automatiquement les listes de vocabulaires des mots simples reconnus, des mots composés reconnus et des mots inconnus des dictionnaires d'Unitex, accessibles depuis les onglets en bas a gauche de l'écran après importation d'un texte dans Unitex.

Introduction à Unitex

La plus simple des fonctionalités d'Unitex permet d'effectuer une recherche textuelle avec un "patron à trous" (expression régulière).



En activant l'item de menu : Text/Locate Pattern ...

- Opérateur de concaténation (le point)

```
le <A> chien
(le <A>)chien
le.<A>chien
(le).<A> chien
(le.(<A>)) (chien)
```

- Opérateur d'alternative (le plus),

```
(je+tu+il+elle+nous+vous+ils+elles)<V>
(le brave garçon)+(le garçon)
```

- Opérateur de répétition (étoile de Kleen), de 0 à N occurences.

```
le brave* garçon
```

Masques lexicaux

- <^> saut de ligne
- <PNC> ponctuation
- <MOT> n'importe quel mot composé de lettres
- <NB> n'importe quel mot composé de chiffres
- les mots du dictionnaire <DIC>
- <A:mp:f> les adjectifs masculins pluriels ou féminins.
- <N~Hum+z1> les noms (N) qui ne référent pas à des humains (~Hum) et qui sont de niveau de langue générale (+z1)

Automates d'états finis déterministes

Qu'est-ce qu'un automate d'états fini déterministe ?
Un automate d'états fini déterministe est une autre représentation d'une expression régulière.
Un automate d'états fini déterministe est défini par :

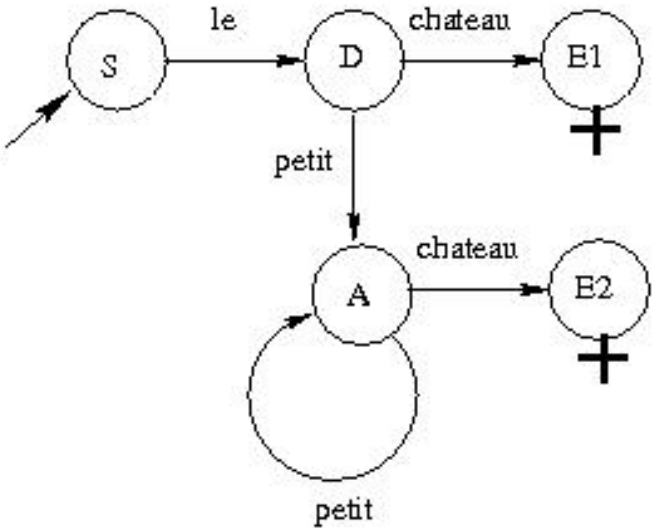
1. un ensemble fini d'états
2. un ensemble fini de transitions (arcs reliants les états)
3. un alphabet d'entrée (liste finie de symboles d'entrée)
4. un état initial
5. un ou plusieurs états finaux.

L'automate est déterministe si dans un état donné, pour un symbole d'entrée donné, on ne peut emprunter qu'une seule transition.

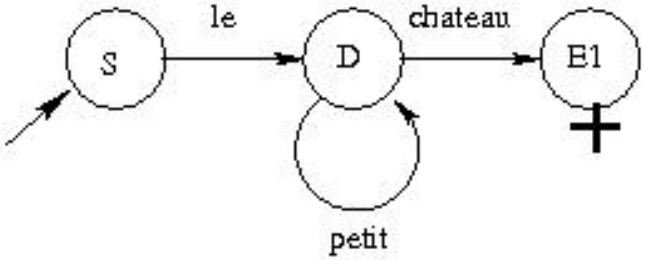
Automates d'états finis déterministes

le chateau
le petit chateau
le petit petit chateau
le petit petit petit chateau
le petit petit petit.....petit chateau

automate non minimal



automate minimal



Automates d'états finis déterministes

Un automate d'états fini déterministe est minimal, s'il n'existe aucun autre automate d'états fini déterministe avec moins d'états qui reconnaisse le même langage (automate équivalent).
Un automate d'état fini déterministe minimal n'a pas de transition vide (souvent notées ϵ).
Un automate déterministe fonctionne en "accepteur" (reconnaissance).

Automates d'états finis déterministes (suite)

Un automate d'états fini déterministe est minimal, s'il n'existe aucun autre automate d'états fini déterministe avec moins d'états qui reconnaisse le même langage (automate équivalent).
Un automate d'état fini déterministe minimal n'a pas de transition vide (souvent notées ϵ).
Un automate déterministe fonctionne en "accepteur" (reconnaissance).
Pour le faire fonctionner en "traducteur" c'est à dire en transducteur, (en anglais *transducer*), il suffit d'ajouter des symboles de sortie sur les transitions.