Bases formelles du TAL : Mini-Projet « Automates »
Distribué le 12 février 2008,
Retour le 8 avril 2008, Soutenance le 15 avril 2008

Objectifs

L'objectif de ce projet est de permettre une recherche dans des textes en décrivant le motif recherché par le biais d'automates.

Plus précisément, il s'agit de réaliser une application capable
– de lire dans un fichier une ou plusieurs descriptions d'automates
– de manipuler ces automates : complétion, déterminisation, suppression des epsilon-transitions, réunion.
– de trouver dans un texte toutes les occurrences d'un mot reconnu par un automate,
– et d'afficher ces occurrences avec un contexte déterminé (une ligne par exemple)
Par ailleurs, pour tester de façon réaliste ce programme, on demande de choisir un phénomène linguistique qui peut être décrit par une série d'automates, et de mener une petite étude linguistique pour définir la série d'automates appropriée. Parmi les phénomènes que l'on peut proposer (d'autres peuvent être envisagés après accord de l'enseignant) :
– inversion locative du sujet (*devant lui se tenait Jean...*)
– adverbiaux temporels (*la plupart du temps*, *mardi à huit heures...*)
– déterminants complexes (*une majorité de*, *entre trois et cinq...*)
– constructions à verbe support (*faire attention*, *prendre un rendez-vous...*)
– formes composées et surcomposées des temps verbaux (*Quand Panturle a eu labouré son champ, il a déjeuné*)
– les compléments d'agent dans les phrases passives (entouré *d'amis*, assassiné *par Brutus...*)
– ...

Recherche dans un texte

On supposera que l'on dispose de textes étiquetés, qui prennent la forme illustrée par l'exemple suivant :

```
On:CL3ms
s':CL3ms
occupe:VP3s
de:P
l':Dfs
entreprise:NCfs
depuis:P
sa:Dfs
naissance:NCfs
jusqu'à:P
sa:Dfs
mort:NCfs
"
,
résume:VP3s
Me:NCms
Jean-Michel:NPms
Lepretre:NPms
,
associé:VKms
de:P
le:Dms
cabinet:NCms
Rambaud-Martel:NPms
.
```

Les motifs recherchés peuvent aussi bien concerner un mot (ou une partie de mot) qu'une étiquette (ou une partie d'étiquette). Il faut donc prévoir une notation permettant de distinguer les deux cas.
Le jeu d'étiquette est décrit à la page
`http://www.linguist.jussieu.fr/~amsili/Ress/jeuEtiquettes.txt`

Format des automates

Le format des fichiers décrivant les automates est libre, mais on peut s'inspirer de l'exemple suivant, où la description donne toutes les informations nécessaires :

```
1 a 2
2 b 2
2 b 5
1 a 3
3 c 2
3 c 4
3 a 3
4 b 3
4 a 5
5
```