Projet 3 : Désambiguation par automates et extraction de catégories à partir d'un texte étiqueté

Il s'agit, à partir d'un texte découpé en mots, chacun étiqueté avec toutes ses catégories grammaticales possibles, de lever certaines ambiguïtés non pertinentes.
On s'attachera à l'extraction d'une catégorie donnée, par exemple « Verbe conjugué ». On extraira du texte une première liste de mots qui peuvent être des verbes conjugués, puis, après désambiguation, une deuxième liste comprenant seulement les mots qui sont effectivement des verbes conjugués dans le texte.
Il est conseillé de procéder comme suit :

1. Étude linguistique

Relever les contextes locaux (catégories précédentes, catégories suivantes) qui permettent de choisir la catégorie retenue pour un mot donné en français.
Par exemple, si X est marqué N ou V, Det + X $\Rightarrow$ X = N.
Etablir ainsi des séquences de 2 ou 3 catégories impossibles en français.
On aura intérêt à subdiviser certaines catégories du DELAF, par exemple mettre à part parmi les pronoms, les formes clitiques (*je*, *tu*, *il(s)*, *elle(s)*, *nous*, *vous*, *se*, *on*, *ce*, *y*, *en*, *le*, *la*, *les*, *lui*, *leur*).
Attention au cas où les mots suivants ou précédents ont plusieurs catégories possibles.

2. Programmation et implémentation

On conseille d'implémenter les règles dégagées sous forme de transducteurs.
Choix du langage : C ou Pascal.
On fournit deux texte étiquetés (d'une dizaine de pages) à l'aide du DELAF, qui serviront de base à l'élaboration des règles de désambiguation. Le programme devra être testé sur ces deux textes, ainsi que sur un troisième qui n'aura pas été étudié a priori.

Rappel des notations du DELAF :

V : verbe	Pour les verbes :
N : nom	P : présent indicatif
DET : déterminant (article)	S : subjonctif présent
DETP : déterminant possessif	Y : impératif
DETQ : déterminant numéral	K : participe passé
INTJ : interjection	F : futur
CNJS : conjonction de subordination	W : infinitif
CNJC : conjonction de coordination	I : imparfait
A : adjectif	C : conditionnel
PRON : pronom	G : participe présent
ADV : adverbe	T : subjonctif imparfait
Xin : préfixe	Inv : invariable
	1, 2 ou 3 : personne
	m, f, s, p : genre, nombre

Exemple Forme générale : *mot lemme cat:morpho ...*

```
vient  venir V33:P3s
livre  livre N1:ms  livre N21:fs  livrer V3:P1s:P3s:S1s:S3s:Y2s
```

Références bibliographiques

Max Silberztein *Analyse lexicale et disctionnaires électroniques*, Masson, 1993

Langue française Numéro spécial de la revue *Langue française*, 1990 (articles de B. Courtois et de M. Silberztein)

Catherine Fuchs Linguistique et traitements automatiques des langues, Hachette U, 1993 (chapitre morphologie)

D. Perrin , « Automates et algorithmes sur les mots », *Annales des Télécoms*, 44:1-2, 1989.

Max Silberztein « A New Approach to Tagging », *Journal of French Language Studies*, CUP, 1991