Données linguistiques et corpus

Cours VI

Grégoire Winterstein
Laboratoire Structures Formelles du Langage, Université Paris 8

gregoire.winterstein@linguist.jussieu.fr

Université Paris Sorbonne

Détails pratiques

Cours

- Le Jeudi, **10h30-12h30**, salle 220 (Serpente)
- **Séances prévues** :

~~16 février~~	~~23 février~~	~~8 mars~~	~~15 mars~~
~~22 mars~~	**29 mars**	12 avril	**03 mai**

- Site web :
  `http://gregoire.winterstein.free.fr/Ens/DonneesLing/`

Évaluation

- Un examen sur table pendant la séance du 3 mai.

Thèmes abordés dans le cours

- Quelles données pour la linguistique ? ✓
- La recherche par expressions régulières. ✓
- Les formats d'encodage des textes. ✓
- Étude d'outils particuliers :
  - TLFi✓
  - Frantext ✓
  - **French Treebank**
- **L'approche expérimentale :**
  - Mise en place d'une expérience.
  - Discussion des résultats.

Projet d'expérience

En prévision de la mise en place d'une expérience :

- Réfléchir à des phénomènes linguistiques que vous désirez tester.
- Construire des paires minimales mettant en jeu les phénomènes en question.

Contenu

Le French TreeBank

- Corpus journalistique en français : environ 1 million de mots, extraits d'articles du journal *Le Monde* de 1989 à 1993.
- Corpus **annoté** en **constituants** et en **fonctions**
- Description détaillée et liens vers la documentation : `http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php`
- Il existe des corpus arborés pour la plupart des langues documentées. Un exemple célèbre : `http://www.ldc.upenn.edu/ldc/online/treebank/` (*Penn Treebank* consultable en ligne)

Consulter le corpus

- Pour consulter et chercher des structures dans le corpus on utilise l'outil `Tregex` : `http://nlp.stanford.edu/software/tregex.shtml`
- `Tregex` est un utilitaire permettant de faire des recherches sur un corpus arboré avec des possibilités analogues à celles offertes par les expressions régulières.
- On utilise une version du FTB au format adapté pour une utilisation avec Tregex (voir page du cours).
- Démarche :
  1. Télécharger l'outil depuis l'adresse suivante : `http://gregoire.winterstein.free.fr/Ens/DonneesLing/tregex.zip`
  2. Extraire les fichiers
  3. Télécharger le corpus du FTB (cf. instructions en cours)
  4. Double-cliquer sur le fichier `stanford-tregex-2012-03-09.jar` pour lancer `Tregex`

Aperçu de la version pour tregex

```
(TOP (VN (CL On) (V devrait) ) (VPinf (VN (CL y) (V
voir) ) (NP (D un) (N gage) (AP (A précieux) ) (PP (P
contre) (NP (D la) (N résurgence) (PP (P d') (NP (D un)
(PREF super-) (N Etat) (AP (A allemand) ) ) ) ) ) ) )
(PONCT .) )
```

Interface Tregex

- Charger le fichier `treebank.tgrep.txt` depuis le menu `File`→`Load Trees`
- Entrer sa requête dans le cadre `Pattern`
- Cliquer sur `Search`
- Les phrases comportant la séquence recherchée apparaissent dans le cadre de droite.
- En cliquant sur une des phrases résultat, sa structure s'affiche dans le cadre du bas.

Utiliser Tregex - éléments de syntaxe

`A,B...` représentent des étiquettes utilisées dans l'annotation (se reporter au guide d'annotation du FTB pour les détails).

- `A << B` : A domine B
- `A < B` : A domine immédiatement B
- `A $ B` : A est un noeud soeur de B
- `A .. B` / `A , , B` : A précède B / A suit B
- `A <, B` : B est le premier fils de A
- Groupements relatifs à l'élément le plus à gauche, gestion des groupes de catégories par parenthèses :
  - `S < VP < NP` : un S qui domine à la fois un VP et un NP
  - `S < (VP < NP)` : un S qui domine un VP qui domine un NP
- Combinaison de relations :
  - `&, |` : conjonction, disjonction
  - `?, !` : optionalité, négation
  - Regroupement de relations avec les `[]` : `NP [< NN | < NNS] & > S`
- Énormément d'autres possibilités, se référer à la documentation `README-tregex.txt`

Recherches dans le corpus : exercices (I)

1. Rechercher les SN sans nom dont le pivot est un adjectif (p.ex. *les plus grands*)
2. Rechercher les séquences verbales composées de trois pronoms clitiques adjacents
3. Rechercher les noyaux verbaux composés de deux verbes séparés par un adverbe
4. Extraire les phrases dans lesquelles l'adverbe **pas** apparaît après un auxiliaire de temps à l'infinitif.
5. Combien de fois la forme **soit** est-elle catégorisée comme conjonction de coordination ? En vous basant sur les 10 premiers résultats, combien d'emplois de cette conjonction distinguez-vous ?

Recherches dans le corpus : exercices (II)

1. Rechercher les groupes prépositionnels figés catégorisés comme adverbes.
2. Rechercher les emplois de la séquence **ainsi que** comme coordonnant complexe.
3. Quel est le pourcentage de phrases (graphiques) dépourvues de tout verbe ?

Jugements linguistiques

- Comment recueillir correctement des jugements linguistiques ?
- Rappel de la démarche :
  - Introspection du linguiste
  - Discussions informelles
  - Questionnaire
    - Identifier les biais.
    - Paires minimales qui mettent en jeu les paramètres à tester dans toutes les combinaisons pertinentes.
  - Expérience formalisée
  - Contrôle statistique des sujets : nombre suffisant pour
    1. faire des statistiques fiables
    2. exclure les sujet aberrants

Préparer son expérience

- Une expérience doit venir **valider** des hypothèses.
- Il faut les formuler **avant** de lancer son expérience et construire l'expérience en conséquence.
- « Chercher » des conclusions dans les résultats d'une expérience n'est pas acceptable :
  - Les conclusions risquent de « coller » aux données particulières d'une expérience
  - On peut se servir d'observations pour formuler de nouvelles hypothèses qu'il est nécessaire de tester avec une nouvelle expérience.