Données linguistiques et corpus

Cours V

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 lors de 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

« *Ne. . . pas* » vs. « *ne. . . point* »

- Corpus : « *La Princesse de Clèves* », ouvrage de 1678
- **Question :** peut-on prédire l'utilisation de *ne. . . point* plutôt que *ne. . . pas* ?
- Comment procéder ?
  - Interroger ses intuitions.
  - Consulter les grammaires.
  - Vérifier sur corpus.

Source 1

- Au XVII^{e} siècle **ne** suffit encore à la négation, mais est de plus en plus systématiquement associé à des termes de "renforcement" : **pas**, **point**.
- Vaugelas (1697) :

> *Il est difficile de donner des règles pour le choix entre "pas" ou "point" : il faut l'apprendre de l'usage et se souvenir que* ***"point" nie bien plus fortement que "pas"****.*

- "Point" jamais devant les noms, toujours suivi de "de" mais sans article défini.

(1)
a. #Il n'a point de l'argent.
b. #Il n'y a point moyen. (faute relevée à la cour)
c. Il n'y a point de moyen.
d. Il n'y a pas moyen.

⇒ suggère que les deux peuvent être en distribution en partie complémentaire (définie par l'usage).

Source 2

Dictionnaire Critique de la Langue Française, Jean-François Féraud (édition de 1788, article **Pâs**)
Voici pourtant quelques règles que donne là-dessus l'Ab. REGNIER.

1. Dans les phrâses de pure négation et de pure prohibition, on se sert plus ordinairement de "pas" que de "point". 'Il "ne" veut "pas" ; il ne prétend "pas". "Ne" le faites, "ne" le croyez "pas".
2. Dans les interrogations, "point" marque un doute, et "pas" une croyance positive : '"Ne" l'avez-vous "point" vu ? "Ne" l'avez-vous pas "vu" ?
3. On peut se servir "de point" à la place de "non", pour répondre négativement à une interrogation : "pas" n'est alors de nul usage. 'En est-il d'acord ? "point" : "point du tout".
4. "Point" ne sympathise pas avec "plus". 'Il "n'"y a "point plus" de réalité dans les aûtres destructions et revivifications que les Alchimistes font soner si haut. "Hist. du Ciel". On doit dire : "il n'y a pas plus", etc. Voy. à sa place, dans l'ordre alphabétique, POINT, particule négative.

Récapitulatif

1. **Point** s'il est suivi d'un nom, doit l'introduire par **de** et sans article défini (au contraire de **pas**).

   (2)
   a. *point N
   b. pas N
   c. *point de Det N
   d. pas de Det N
   e. ?point Det N
   f. pas Det N
   g. point de N
   h. pas de N

2. **Point** ne se combine pas avec **plus**
3. Dans les impératifs : usage préférentiel de **pas**

$\Rightarrow$ Tester ces différentes hypothèses avec Frantext

Tester avec Frantext

1. Comparer le nombre de négations avec **pas** et celles avec **point**.
2. Comparer le nombre de négations avec **pas** et celles avec **point** sur un même auxiliaire.
3. Tester les combinaisons de (2) qui sont testables
4. Tester la combinatoire de **plus** avec **pas** et **point**
5. Tester l'emploi de **point** dans des impératifs (trouver une forme caractéristique).

Une étude des verbes

- On peut émettre l'hypothèse que certains verbes préfèrent **point** à **pas**
- Si on relève le nombre de fois qu'un verbe utilise **pas** ou **point** pour sa négation on peut ensuite comparer les fréquences d'emploi.
- Exemples :
  - **craindre, douter, demander, parler** semblent préférer **point**.
    - Avec ces verbes si **pas** est employé c'est généralement à un mode différent de l'indicatif.
  - **croire** n'emploie jamais **point**
  - . . .

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)

Aperçu

```
 <SENT>
<PP fct="MOD">Au <NP>début</NP></PP>,
<VN fct="SUJ">on ramassait</VN>
<VPinf fct="OBJ">
<PP fct="DE-OBJ">de <NP>quoi</NP></PP>
<VN>remplir</VN>
<NP fct="OBJ">quinze sacs_poubelle</NP>
</VPinf>,
<Sint>
<VN>indique</VN>
<NP fct="SUJ">Roger, <NP>ouvrier <PP>? <NP>la
régie</NP></PP></NP></NP>
</Sint>.
</SENT>
```

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`

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 ?