Données linguistiques et corpus

Cours IV

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	**~~5 avril~~ ⇒19 avril ?**	12 avril

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

Évaluation

- Un examen sur table (séance à préciser).

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

Expressions de séquence

Suite de **descriptions** de mots en partie inspirée de la syntaxe des expressions régulières et réminiscente des codes du TLFi :

- Graphie exacte d'un mot, p.ex. `mangeront`
- Disjonction `(X|Y)` : tout élément qui vérifie la description *X* ou *Y*
- `&?xxx` : un élément optionnel qui vérifie la description *xxx*
- `&cxxx` : toutes les formes de l'infinitif *xxx*
- `&mxxx` : toutes les formes de l'adjectif ou substantif *xxx*
- `&lxxx` : tous les membres de la liste nommée *xxx* (créée comme dans le TLFi, cf. barre de gauche)
- `&q(`n_1`,`n_2`)` : une suite de mots quelconque dont le nombre d'éléments est compris entre n_1 et n_2 (raccourci : `&q=&q(1,1)`)
- `^xxx` : tout élément qui ne correspond pas à la description *xxx*
- `&e(xxx)` : tout élément qui vérifie les informations de catégorisation spécifiée par *xxx* (uniquement pour la recherche dans Frantext catégorisé), cf. infra

Informations de catégorisation

Utilisation de la description `&e()` :

- `&e(g=X Y)` : tout élément qui possède la catégorie X ou Y
- `&e(c=X)` : tout élément catégorisé qui vérifie l'expression de séquence X
- `&e(g!=X Y)` : tout élément qui ne possède ni la catégorie X ni la catégorie Y (négation identique pour `&e(c!=X)`).
- Combinaisons possibles :
  `&e(c=(&mdemi &q | &q &mdemi) c!=&mheure g=S)`

Codes grammaticaux courants :

`A`	Adjectif	`Inf`	Infinitif
`Adv`	Adverbe	`P`	Pronom
`Cc`	Conj. coordination	`Pp`	Préposition
`D`	Déterminant	`S`	Substantif
		`V`	Verbe

Voir la liste complète dans l'aide de Frantext.

Aller plus loin : création de “grammaires”

- Moyen de formuler des expressions de recherche flexibles et puissantes.
- Lien ténu avec la conception de grammaire scolaire.
- Inspirées des grammaires de réécriture :

$$
\begin{array}{lcl}
S & \rightarrow & SN\ SV \\
SV & \rightarrow & Vi \\
 & | & Vt\ SN \\
SN & \rightarrow & Det\ N
\end{array}
$$

etc.

Définitions de grammaires dans Frantext

- Grammaire = Ensemble de règles composées d'un **nom** et d'un **corps**.
- Le corps d'une règle correspond à une suite d'expressions de séquences.
- On fait appel à une grammaire dans une expression de séquence via le code `&rxxx,yyy` (=appel de la règle *xxx* dans la grammaire *yyy*, ce dernier paramètre est omis si la grammaire est la grammaire courante).

Grammaires : un exemple pour le GN

```
gp :
&e(g=Pp) (&e(g=Np)|&e(g=D) &e(g=S))
ga :
&?(&e(g=Adv)) &e(g=A) &?(&rgp)
gncommun :
&e(g=D) &?(&rga) &e(g=S) &?(&rga) &?(&rgp)
gnpropre :
&?(&e(g=D)) &e(g=Np)
gncomplet :
(&rgncommun|&rgnpropre)
```

- Suggestions d'améliorations ?

⇒ récursivité de la règle du PP :

```
gp :
&e(g=Pp) (&e(g=Np)|&e(g=D) &e(g=S))
```

- **Mais...** la récursivité est interdite dans Frantext.

Grammaires paramétrées

- Les grammaires sont paramétrables, on peut écrire les règles de manière à pouvoir en spécifier une partie au moment de leur appel :
- `gncommun` :
  `&e(g=D) &?(&rga) &e(g=S c=&m&1) &?(&rga) &?(&rgp)`
- Appel avec un paramètre :
  `&rgncommun(chien),G`
  appelle la règle paramétrée `gncommun` de la grammaire `G` avec la valeur 'chien' pour le premier paramètre.
  Résultat : ne recherche que les groupes nominaux centrés sur les chiens.

Étude de voisinage

- Permet de déterminer les éléments lexicaux qui apparaissent le plus fréquemment autour d'un item (ou d'une liste d'items).
- Possibilité de définir le voisinage :
  - En nombre de phrases autour de la phrase contenant l'élément pivot (= l'élément recherché).
  - En nombre de mots autour du pivot.
- **Résultat :** liste de mots assortis de leur nombre d'occurrences dans le voisinage spécifié.
- **Attention :** les pivots sont inclus dans la liste des résultats (ils sont dans leur propre voisinage).

Calculs de fréquence

- Permet d'étudier le nombre d'occurrences d'un terme (ou d'une liste de termes) dans le corpus de travail.
- Possibilité de :
  - Étudier l'évolution de la fréquence selon un paramètre du corpus (date, auteur. . .)
  - Spécifier des mots par expressions régulières : *Fréquence des mots du corpus de travail*
- **Résultat :** liste des mots assortis de leur fréquence dans le corpus.

Exercices (I)

1. Quelle est la première attestation de **morphème** dans le corpus ?
2. Georges Perec utilise-t-il le terme **ordinateur** ?
3. Au présent de l'indicatif, quelle est la forme la plus fréquente du verbe **subir** ?
4. Déterminez si l'inversion du sujet est possible après **aussitôt que**.
5. Peut-on insérer un élément entre **tandis** et **que** dans la « conjonction de subordination » **tandis que** ?
6. Trouvez tous les exemples dus à Balzac d'utilisation d'un substantif qui est également une forme du verbe **voir**.

Exercices (II)

1. L'expression **faire amende honorable** s'emploie-t-elle au passif ?
2. La forme **sue** est-elle plus souvent une forme du verbe **savoir** ou du verbe **suer** ?
3. De quand date le substantif **déterminant** pris dans son sens linguistique ?
4. A quelle fréquence rencontre-t-on la négation exprimée par **ne** sans **pas** (par rapport à la négation avec **pas**) pendant la première moitié du XX^{e} siècle ?
5. Comparez la fréquence des verbes du premier groupe, à la première et à la troisième personnes du pluriel, au passé simple, au XIX^{e} et au XX^{e} siècles.