Données linguistiques et corpus

Cours II

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	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.
- ...

Contenu

Rappel : la question de régularité verbale

Deux options :

1. Régulier = le plus représenté.
   - Le patron *base*, *base+ed*, *base+ed* est effectivement le plus représenté en anglais.
2. Régulier = être capable de conjuguer un verbe sans avoir besoin de l'apprendre par cœur.

   $\Rightarrow$ Un verbe irrégulier doit être appris par cœur.

Alternative

Hypothèses pour le français :

1. La régularité est distincte de la fréquence, et seul le premier groupe est régulier.

   $\Rightarrow$ les locuteurs ne savent conjuguer intuitivement que les verbes du premier groupe.

2. La régularité est distincte de la fréquence, et les premier et deuxième groupes sont réguliers.

   $\Rightarrow$ les locuteurs savent conjuguer intuitivement les verbes des premiers et deuxième groupes.

3. La régularité est la manifestation directe de la fréquence.

   $\Rightarrow$ il y a un gradient de facilité à conjuguer les verbes des premier, deuxième et troisième groupe.

Comment trancher ?

- Enregistrer enfants, apprenants, adultes compétents : interminable pour obtenir suffisamment de données.
- Examiner les lexèmes nouveaux : les facteurs mis en cause dans la lexicalisation de lexèmes sont complexes.
- Une bonne solution : le **wug-test**

THIS IS A WUG

NOW THERE IS ANOTHER ONE.

THERE ARE TWO OF THEM.

THERE ARE TWO _____.

Un bon *wug*

- Créer de toutes pièces un nouveau verbe (un **logatome**).
- Soumettre ces verbes aux locuteurs pour qu'ils les conjuguent.
- Un bon logatome :
  - sonne "français", cf. **shmurdzer**
  - n'est pas proche d'un item existant, cf. **boulanger**.
- Quelles formes tester ?
  - Si on part de l'infinitif, il est facile de conjuguer le présent :
    bruglir→ *je bruglis* / *brugler*→*je brugle*
  - L'inverse est plus intéressant :
    je bruglis→*bruglier* (cf. *publier*) ou *bruglir* (cf. *faiblir*) ?

Conditions de test

- Introspection du linguiste
- Discussions informelles
- Questionnaire :

  (1) Tous les jours je bruglis/brugle. C'est Pierre qui m'a appris à...

- Expérience formalisée :
  - Contrôle statistique des sujets : nombre suffisant pour
    1. faire des statistiques fiables
    2. exclure les sujet aberrants
  - Mesurer des paramètres objectifs :
    - temps de réponse
    - mouvements de l'œil
    - signaux cérébraux

Quels biais ?

- Quelle est la tâche exactement demandée aux sujets ?
  - Conjuguer le verbe comme s'il était nouveau, inexistant ?
  - Conjuguer le verbe comme s'il était déjà dans le dictionnaire ?
- Honnêteté vis-à-vis des sujets : il est préférable qu'ils ne soient pas au courant du but de l'expérience.
- Identification de la tâche : des items de contrôle sont nécessaires pour vérifier que les sujets effectuent bien la tâche qui est attendue de leur part.

Bilan

- Pour certains problèmes une expérimentation psycholinguistique semble être le seul moyen d'accéder aux données pertinentes.
- Néanmoins, c'est un travail lent et complexe, notamment du fait du grand nombre de paramètres qui rentrent en jeu.
- La question de la nature de la régularité est loin d'être réglée :
  - En anglais il existe beaucoup de littérature, aujourd'hui inconclusive.
  - Sur le français et les autres langues à conjugaison complexe il n'existe que très peu d'études.

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.

Corpus oraux

- http://www.loria.fr/projets/asila/corpus_en_ligne.html
- http://www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/Inventaire.pd

Exemple de contenu

```
<u id="mf4u209" who="E">
et puis le <pause dur="short"/>ceux qui ont pris leur parachute par
deux ficelles ou par une ficelle </u>
<u id="mf4u210" who="N">
ouais parce que moi <pause dur="short"/>j'avais une ficelle <pause
dur="short"/>il s'est pas ouvert</u>
<u id="mf4u211" who="E">
ouais j'en ai pris deux ficelles <pause dur="short"/>il s'est super
bien ouvert</u>
```

Expressions régulières

- À la base un moyen de caractériser un langage du point de vue formel (i.e. comme un ensemble de mots obéissant à certaines contraintes de forme).
- Aujourd'hui : outil de recherche dans un document texte.
- Au lieu de chercher une séquence de lettres données, on cherche toutes les séquences qui correspondent à un **motif**.

Construction d'un motif

- Une suite de caractères non-spéciaux est un motif (attention aux majuscules) : 'partir' reconnaît *partir* (mais pas *Partir*).
- Caractères spéciaux :
  - . remplace tout caractère. 'cha.on' : *chapon, chaton...*
  - caractères de répétition $*, +, ?$
    - 'livre.*' : *livre, livres, livreur, livrer...*
    - '.+ger' : *manger, loger, gruger...*
    - 'livres?' : *livre, livres*
  - [] spécifie un ensemble de caractères : 'cha[pt]*on*' : *chapon, chaton*
  - [^] spécifie le complément d'un ensemble de caractère, 'cha[^pt]*on*' : *charon, chalon*
  - On peut spécifier une plage de caractères : [a-z]= tout caractère de 'a' à 'z'.
  - | marque une disjonction. cent(re|er) : *center, centre*
  - Chercher un caractère spécial ?
    $\Rightarrow$ caractère d'échappement '\'
  - \b indique une frontière de mot
  - Beaucoup d'autres options selon les outils (répétition contrôlée, groupe de caractères etc.)

Quels outils pour utiliser les expressions régulières ?

- Word et OpenOffice (“utiliser les caractères génériques”/ “Expressions régulières”)
- Éditeurs de texte (emacs, vi, notepad++. . .)
- Google (au moins *, et via l’emploi de ")
- Langages de programmation accessibles : python, perl. . .

TP

Avec l'outil disponible sur `http://www.gskinner.com/RegExr/` :

- Recopier le contenu du fichier `pendule.txt` placé sur le site du cours.
- Y effectuer les recherches suivantes :
  - repérer le nombre d'occurrences du mot *gravure*
  - repérer les occurrences du mot `métal` au singulier et au pluriel
  - repérer le nombre d'occurrence des pronoms de troisième personne.
  - repérer tous les mots qui finissent par `ique` ou par `sque` (éventuellement au pluriel)
  - repérer toutes les séquences constituées de `le` ou `la` suivi d'un mot au singulier
- Trouver une expression régulière qui reconnaisse les adresses e-mail bien formées et la tester sur des exemples comme :
  - bob@truc.fr
  - bob.jack@test.co.uk
  - * bob@site
  - * bob@mail@domain.com
  - * bob.mail.fr

Remplacements

- Outre chercher des chaînes de caractères dans un texte, les expressions régulières peuvent aussi servir à **remplacer** ces chaînes par d'autres.
- **Ex.** remplacer toutes les conjugaisons de première personne du pluriel par la deuxième personne du pluriel :
  1. Comment capturer ces éléments ?
     - terminaisons en **...ons** : `\b\S+ons\b`
  2. Quoi remplacer ? Quoi conserver ?
     - Pour remplacer : `$1` réfère au "groupe" n° 1 (marqué par des parenthèses dans le motif).
     - Essayer de remplacer '`\b(\S+)ons\b`' par '`$1ez`'.
     - Problème avec *mangeons* → *mangeez*.
     - **Recherche** : '`\b(\S+[^e])e?ons\b`'
     - **Remplacement** : `$1ez`
- Pour plus d'information consulter internet.

Le TLFi

- **T** résor de la **L** angue **F** rançaise **i** nformatisé
- Disponible à l'adresse : `http://atilf.atilf.fr/tlf.htm` (ou sur CD-Rom)
- Dictionnaire ancien : XIX^{e} et XX^{e} siècle, dernier supplément en date de 1994.
- Pas toujours idéal :
  - on y trouve des mots archaïques (p.ex. *azamoglan*).
  - aucun terme récent ne s'y trouve (p.ex. *navigateur, cédérom*)
- Ressource structurée, avec des possibilités de recherche avancées.

Recherche de base

Trois possibilités :

1. Recherche directe d'un mot (éventuellement fléchi)
2. Liste défilante (ne concerne que les entrées principales)
3. Saisie "phonétique" (*B-É*)

- Largement suffisant pour chercher une définition donnée, pas assez puissant pour des recherches lexicographiques un peu poussées.

Recherche avancée

2 onglets (haut de la page)

Recherche assistée : 5 façons de définir une recherche.

- Spécialisation par discipline, code grammatical, emploi (invariable, argotique. . .)
- Possibilité des **codes de contenu** dans le cadre 1 (cf. infra)
- Contraintes liées aux éléments de l'entrée (définition, langue d'emprunt. . .)

Recherche complexe : jusqu'à 6 contraintes liées sur une même recherche

- **Exemple** : verbes à usage péjoratif empruntés à l'italien et y repérer la définition appropriée.

Codes de contenu

- `&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* (cf. infra)
- `&q` : n'importe quel mot
- `^xxx` : toute séquence qui ne correspond pas à la séquence *xxx*
- `&nxxx` : toute séquence qui ne contient pas *xxx*
- `&dn xxx` : signifie que le contenu *xxx* doit se trouver à *n* mots au plus en partant du début de l'objet étudié.
- `&fn xxx` : signifie que le contenu *xxx* doit se trouver à *n* mots au plus en partant de la fin de l'objet étudié.
- `|` relie des contraintes multiples (**attention** $\neq$ regexp)

```
&cmanger &q &mpain
^&cruer dans les brancards
```

Listes de mots

- Utile pour créer des listes de mots possédant des propriétés similaires.
- Ces listes sont ensuite réutilisables dans les recherches complexes (cf. code `&l`).
- À la création d'une liste, son contenu s'affiche dans le cadre 1.
- Utilisation des **expressions régulières** (cadre 4) :
  - Caractère joker : .
  - Répétition : $*$, $+$, ? et \1 (p.ex. `(co)\1` renvoie *coco*)
  - Disjonction : |
  - Ensemble de caractères : []
  - Négation : [$^{\wedge}$*xyz*]

Exercices

1. trouver tous les verbes empruntés à l'arabe
2. trouver tous les noms qui riment avec *rieur*
3. trouver tous les exemples d'emploi du nom *agrégation* au sens du concours de l'agrégation
4. trouver tous les mots en *ing* empruntés à l'anglais
5. trouver toutes les expressions figées basées sur le verbe *attraper*, du type *attraper froid*
6. trouver tous les mots échos (p.ex. *bonbon*, *chouchou*...)
7. trouver tous les palindromes qui contiennent entre 6 et 8 lettres