Données linguistiques et corpus

Cours I

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 (court) dossier à rendre ou un examen sur table...

Thèmes abordés dans le cours

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

Contenu

La linguistique comme science

- Linguistique : étude (scientifique) du langage humain.
- Le linguiste émet des **hypothèses falsifiables** sur le langage qu'il confronte ensuite à la réalité du langage.

$\Rightarrow$ Définir une notion d'**observable** pour le linguiste.

Exemple non-linguistique

- La mécanique étudie le mouvement des corps.
- Les observables : positions dans l'espace à des instants donnés.
- Deux théories : mécanique newtonienne vs. relativité générale.
- Cas test : éclipse solaire et courbure des rayons lumineux. Chacune des théories fait une prédiction différente.
- Une mesure est prise : on tranche en faveur de la relativité générale.

Quelles données pour la linguistique ?

Quelques exemples. . .

Phonologie : formes alternatives

(1) a. *[ʒənətələdəmɑ̃dəɣɛpa]
b. [ʒəntəldəmɑ̃dɣɛpa]
c. [ʒəntlədmɑ̃dɣɛpa]
d. [ʃtəldəmɑ̃dɣɛpa]

Morphologie : existence d'une forme

(2) a. institutrice
b. *professrice

Syntaxe : acceptabilité d'une phrase

(3) a. Jean a décidé de lui parler.
b. *A de décidé Jean lui parler.

Quelles données pour la linguistique ? (II)

Sémantique : interprétation d'une expression référentielle.

(4) Paul$_i$ a rencontré Marie$_j$.

a. Il$_i$ a décidé de lui$_j$ parler.
b. *Il$_i$ a décidé de lui$_i$ parler.
c. *Il$_j$ a décidé de lui$_i$ parler.

Pragmatique : adéquation d'un énoncé à un contexte.

(5) ?Le Roi de France est chauve.

(6) a. Paul est grand, mais il est plus petit que Pierre.
b. #Paul est grand, mais il est plus grand que Pierre.

Variation dans le temps (diaphasique)

Amis lecteurs, qui ce livre lisez,
Despouillez vous de toute affection ;
Et, le lisant, ne vous scandalisez :
Il ne contient mal ne infection.
Vray est qu'icy peu de perfection
Vous apprendrez, si non en cas de rire ;
Aultre argument ne peut mon cueur elire,
Voyant le dueil qui vous mine et consomme :
Mieulx est de ris que de larmes escripre,
Pour ce que rire est le propre de l'homme.

Rabelais, Gargantua (1542)

⇒ nécessité de fixer un cadre temporel pour l'étude de la langue.

- **Problème 1 :** certains locuteurs emploient des formes "anciennes" :

(7) a. Je le peux apprécier.
b. Paul ne viendra.

- **Problème 2 :** à date fixe, la **génération** du locuteur importe aussi.

Variation dans l'espace (diatopique)

- Expérience immédiate : les accents régionaux.
- Lexique spécifique à une région : *wassingue*, *chicon*, *schlouk*, *poche*...
- Différences syntaxiques :

(8) a. J'ai fait tomber mon stylo.
b. J'ai tombé le stylo.

(9) a. La mer monte jusqu'à la route.
b. La mer monte jusque la route.

Variation socialement déterminée

(10) a. Paul ne viendra pas.
b. Paul viendra pas.

(11) a. Donne m'en.
b. Donnes-en moi.
c. Donne moi z'en.

(12) a. M'en donne pas.
b. Donne m'en pas.
c. Donnes-en moi pas.
d. Donne moi z'en pas.

(13) a. La voiture de Jean.
b. La voiture à Jean.

- Une description exhaustive de la langue implique de tenir compte de toutes les formes ci-dessus.
- Une langue n'est pas un système linguistique unique et monolithique, mais fonctionne comme une collection de variétés subtilement différentes les unes des autres.
- Les variétés sont socialement marquées : leur usage dépend du contexte social.
- Chaque locuteur maîtrise plusieurs variétés, mais peu les maîtrisent toutes.

Grammaire prescriptive

- La grammaire scolaire présuppose l'existence d'une "bonne" forme.
- Exemples :

(14) a. ✓ Aller chez le coiffeur.
b. × Aller au coiffeur.

(15) a. ✓ Après que je suis parti.
b. × Après que je sois parti.

(16) a. ✓ Aller à vélo.
b. × Aller en vélo.

- Les "mauvais" exemples sont tous massivement employés.
- La prescription rejoint parfois une variation socialement déterminée, mais pas toujours (e.g. **en revanche** vs. **par contre** chez les "grands auteurs").

En résumé

- La collecte de données linguistiques fait face à divers obstacles.
- Cela ne doit pas remettre en cause l'idée qu'il existe des faits stables. Par contre il est impératif de les relativiser à une communauté, une époque, une situation donnés.
- Comme pour toute science empirique il faut se doter d'un appareillage et de méthodes de collecte de données.
- Hypothèses pour avancer :
  1. Le français peut se définir comme un **ensemble d'énoncés possibles**.
  2. Les locuteurs ont une **connaissance implicite** de la langue qui leur permet de produire des énoncés qui appartiennent au français.

L'appareillage linguistique

On considère deux façons d'obtenir des données :

1. L'emploi de **corpus** :
   - Un grand nombre d'énoncés produits par des locuteurs natifs est collecté.
   - Sur la base de *ce qui a été dit* on émet des hypothèses sur *ce qui peut se dire*.
2. Les méthodes dites **expérimentales** :
   - Les locuteurs d'une langue connaissent le système de leur langue : ils peuvent comprendre et produire des énoncés inédits.
   - Cette connaissance est explicitement adressée en proposant des stimuli linguistiques à des locuteurs et leur demandant de porter un jugement sur l'énoncé.
   - Les jugements *conscients* donnent accès à la *compétence inconsciente* des locuteurs.

Corpus ou Jugements ?

- Historiquement, il y a eu une polarisation du débat entre les partisans des deux méthodes au détriment d'une qualité méthodologique.
- Leçon des autres disciplines scientifiques : toute donnée est bonne à prendre.
- Aujourd'hui :
  - Constitution de corpus larges et utilisables (i.e. annotés)
  - Utilisation de protocoles expérimentaux contrôlés et recours aux statistiques pour évaluer les résultats.

Corpus

Corpus : un ensemble **structuré** de textes, idéalement de **grande taille**.

Annotation : souvent, un corpus n'est pas juste une collection de textes mais contient des **annotations** particulières, p.ex.

- parties du discours (FranText)
- structure syntaxique (French TreeBank)
- liages anaphoriques
- indications phonétiques/prosodiques
- ...

Langues : monolingue ou multilingue, dans ce dernier cas :

- Textes différents d'une langue à l'autre mais avec une thématique proche (p.ex. C-ORAL-ROM).
- Textes identiques dans toutes les langues, dans ce cas les textes peuvent être **alignés** (on fait correspondre les pans de texte qui sont des traductions). (p.ex. European Parliament Corpus)

Les écueils des corpus

- Le problème des corrections :

(17) a. Je pense. . . je. . . je cr. . . je crois que. . . que Marie a raison.
b. Je crois que Marie a raison.

- Les erreurs :

(18) a. *Qu'est-ce que vous disez ?
b. Qu'est-ce que vous dites ?

- Les phénomènes rares :

(19)
a. *Paul et moi vivaient dans cette maison.
b. Paul et moi vivions dans cette maison.
c. Dans cette maison ne vivaient que Paul et moi.
d. *Dans cette maison ne vivions que Paul et moi.

La loi de Zipf

- La fréquence d'un mot dans un corpus est inversement proportionnelle au rang qu'il occupe dans la table des fréquences.

⇒ le mot le plus fréquent apparaît deux fois plus que le deuxième mot le plus fréquent, trois fois plus que le troisième plus fréquent etc.

- Sur un corpus d'un million d'occurrences (Brown Corpus), 135 mots suffisent à décrire la moitié du corpus.

⇒ Relativise l'importance d'un "résultat négatif" : la probabilité d'observer un item rare est faible sur un corpus de taille finie.

Des corpus assez grands ?

> *Les cas les plus complexes, qui distinguent les hypothèses, se présentent rarement. On peut facilement vivre toute sa vie sans produire un exemple pertinent montrant qu'on utilise une hypothèse plutôt que l'autre.*
>
> *Chomsky (1980), trad. O. Bonami*

- P.ex : Quelle règle pour l'inversion du sujet dans les questions (en) ?

  (20) Is Paul talking to my father ?

  1. Mettre en tête le premier auxiliaire qui vient ?
  2. Mettre en tête l'auxiliaire de la principale ?

- Pour trancher : recours à des phrases contenant deux auxiliaires, dont un dans le sujet :

  (21) The man Paul is talking to is my father.

  a. Is the man Paul talking to is my father ?
  b. Is the man Paul is talking to my father ?

Des corpus assez grands ? (2)

- Pullum et Scholz (2003) : étudient la structure en question sur le corpus du Wall Street Journal (23000 questions).
- Plus de 1% des questions ont la structure pertinente :

  (22) Is what I'm doing in the shareholder's best interest ?

- Avant l'âge de 3 ans, un enfant entend environ 750000 questions.
- Il a accès a environ 7500 questions avec la bonne structure.
- Ici l'argument de la pauvreté des corpus ne tient pas, mais ce n'est pas toujours vrai (cf. (19)).

Des corpus assez représentatifs ?

- Tension exhaustivité / homogénéité :
  - Si le corpus est très homogène il est difficile de distinguer ce qui relève d'un genre particulier ou du français en général.
  - Si le corpus est très hétérogène encore faut-il savoir décider d'une répartition représentative des genres.
- Cas d'école : les corpus journalistiques (ou littéraires) sont quasi-exempts d'interaction réelles.
- Solution : constituer des corpus nombreux et divers (p.ex. ESTER)

Les écueils des jugements

- Rien ne prouve que la connaissance inconsciente des locuteurs puisse être mobilisée de manière consciente.
- De fait, il est difficile pour certains locuteurs de produire des jugements de grammaticalité.
- Comment juger ces phrases ?

(23) a. L'enfant que l'homme que Marie a croisé a vu a parlé
b. On a perdu le tube dans lequel avait pensé les mélanger notre préparateur les unes avec les autres.

- On peut envisager d'interroger des linguistes, mais leur jugement sera biaisé par leurs connaissances.
- On peut entraîner des non-spécialistes. Mais l'entraînement peut aussi créer un biais.

Un exemple : le *Tequila Test*

Contexte

- Hier soir, il y a eu une fête étudiante regroupant 90 étudiants.
- 30 étudiants ont bu du jus d'orange et rien d'autre.
- 30 étudiants ont bu de la tequila et rien d'autre.
- 30 étudiants ont bu du jus d'orange et de la tequila.

(24) Combien d'étudiants n'ont pas bu seulement de la tequila ?

a. 0
b. 30
c. 60
d. 90

Collecter des jugements de “grammaticalité”

- Quelle que soit la manière dont on pose la question, on observe des différences individuelles dans les jugements des locuteurs.
- La collecte des jugements est sujette à des biais de plusieurs types :
  - Des variations dans les grammaires des sujets
  - Des variations dans les stratégies de réponse
  - Des biais induits par la tâche

Biais liés aux locuteurs

- Les sujets sont plus ou moins habiles à différencier les aspects d'un problème cognitif :
  - p.ex. différence perception visuelle vs. sensorielle.
- Ce facteur influence aussi la capacité des sujets à séparer les facteurs syntaxiques, sémantique etc.
  - p.ex. la phrase suivante est-elle grammaticale ? Si non, pourquoi ?

    (25) Lemmy a résolu tous les problèmes, mais Ritchie quelques-uns.
- Le genre de la personne produit également des différences entre locuteurs (c'est plus probablement un effet social que biologique, cf. les différences de production des voyelles).

Biais liés aux locuteurs (2)

- L'éducation "linguistique" induit également un biais important :
  - Opposition de principe de beaucoup de linguistes : un individu qui a reçu une éducation linguistique est influencé par ses propres conceptions.
  - Position extrême inverse : un entraînement est nécessaire pour produire des jugements fiables.
  - Pas de moyen simple de trancher (p.ex. recours à des techniques d'EEG ou imagerie cérébrale)
  - Une chose sûre : les linguistes et les non-linguistes ne produisent pas les mêmes jugements conscients.
- L'observation s'étend à toute forme d'éducation.
- Le multilinguisme complique encore le problème.

Biais liés à la tâche

- Les locuteurs peuvent comprendre la tâche de manières différentes.
- Lors de jugements graduels, les locuteurs n'ont pas les mêmes repères.

(26) Noter le caractère naturel du texte ci-dessous entre 1 (impossible) et 10 (parfaitement naturel).

a. Marseille est sûre de gagner son match. Bordeaux aussi a des chances de gagner.

- L'ordre de présentation des items de test a une influence : le jugement sur la phrase $n+1$ dépend de la phrase n.
- Répétition : à force de juger, les contrastes s'atténuent.
- Etat mental : en manipulant l'attention des locuteurs on produit des différences de jugement.
- Modalité écrite/orale :
  - Influence évidente du code écrit
  - Normativité variable
  - Ponctuation et intonation

Biais liés à la tâche (2)

- Vitesse de présentation des données.
- Présentation d'un contexte pragmatique.

(27) a. On cherche un acteur pour jouer aux côtés de Pierre. Il ne doit pas être trop petit, mais ne doit pas voler la vedette à Pierre.
b. Paul est grand, mais il est plus grand que Pierre.

- Fréquence des mots a une influence directe.

En résumé

Les linguistes savent bien que les intuitions sur l'acceptabilité des énoncés tendent à être vagues et inconsistantes, dépendant de ce que vous avez mangé au petit-déjeuner et de ce que prédit votre théorie préférée.
O Dahl, 1979 (trad. O. Bonami)

Conclusion

- La conclusion est la même que pour toute science empirique : il existe des biais qu'il est nécessaire de contrôler :
  - en évitant les facteurs de biais identifiés
  - en contrôlant statistiquement les résultats
  - en discutant ses résultats avec la communauté des pairs

$\Rightarrow$ aucun problème pour utiliser des techniques expérimentales en linguistique, tant que les contrôles nécessaires sont en place.

Un cas concret : la conjugaison irrégulière

Bonami & Boyé (2008) « Quels verbes sont réguliers en français ? »

Verbe irrégulier en anglais :

- 5 formes pour un verbe *love*, *loved*, *loved*, *loves*, *loving*
- Focus sur : base, prétérit, participe passé
- Deux règles :
  1. prétérit = base + *ed*
  2. participe passé = prétérit
- Un verbe **irrégulier** est un verbe qui ne respecte pas les deux règles ci-dessus.
- Régulier $\neq$ simple : *cut*, *cut*, *cut* est irrégulier.
- Régulier $\neq$ distinctif : *sink*, *sank*, *sunk* est irrégulier.

Qu'est-ce que l'irrégularité ?

Deux possibilités :

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.

- Chaque option paraît pertinente.
- Peut-on les départager ?

Qu'est-ce que l'irrégularité ? (2)

- Les enfants traitent les irréguliers comme réguliers : *drink*, *drinked*, *drinked*
- Les apprenants étrangers également
- Les locuteurs natifs commettent occasionnellement des erreurs
- Les néologismes sont réguliers : *text*, *texted*, *texted*
- Les emprunts tendent à être réguliers : *sauté*, *sautéed*, *sautéed*
- Les verbes dérivés d'irréguliers sont parfois réguliers : *broadcast*, *broadcasted*, *broadcasted*

Pour le français ?

- 51 formes pour un verbe contre 5 en anglais

$\Rightarrow$ un verbe peut être régulier sur une sous-partie du paradigme

- Il existe des patrons de conjugaison (Bescherelle etc.)
- La notion d'irrégulier n'est pas dans les grammaires "classiques"
- Si on veut l'appliquer :
  1. Premier groupe régulier, le reste irrégulier
  2. Premier et deuxième groupes réguliers, troisième irrégulier
- Le deuxième groupe est aussi "simple" que le premier groupe et le plus fréquent après le premier.

Alternative

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.