Linguistique de Corpus

Elisabeth DELAIS-ROUSSARIE
elisabeth.roussarie@wanadoo.fr

Objectifs du cours

L'objectif de ce cours est triple :

- Expliquer ce qu'est un corpus, en opérant une distinction entre le corpus en linguistique et la linguistique de corpus;
- Apprendre à constituer et à utiliser un corpus : les outils, les standards, etc
- Etablir les faits en linguistique et les corpus

Evaluation

- Un ou deux travaux à effectuer en groupe (nombre sera fixé très rapidement)
- Un devoir sur table (dont la date précise est encore à déterminer)

Planning des séances (1)

Nous aurons huit séances de trois heures réparties comme suit :

Séance 1 : vendredi 17 sept.

Séance 2 : vendredi 1 oct.

Séance 3 : vendredi 15 oct

Séance 4 : vendredi 29 oct.

Séance 5 : vendredi 5 nov.

Séance 6 : vendredi 19 nov.

Séance 7 : vendredi 3 déc. (??)

Séance 8 : vendredi 17 déc. (??).

Planning des séances (2)

- Les dates de décembre restent à confirmer.
- Il se peut que le devoir sur table soit fait pendant la semaine d'examen, donc après les vacances de Noël. La 8ème séance serait donc remplacée par la séance d'examen.

Dans tous les cas, je vous préviendrai par email.

De même, le cours a normalement lieu en salle 065E (Halle aux farines), mais certaines séances pourront peut-être avoir lieu en salle informatique.

Quelques questions d'après les objectifs

L'un des objectifs majeurs de la *linguistique de corpus*, et par voie de conséquence du cours, est d'appréhender une approche en linguistique qui a recours (ou utilise) le corpus pour établir les faits linguistiques.

- Qu'est-ce que la linguistique ? Quelles sont ces objectifs ?
- Qu'est-ce qu'un fait linguistique ?
- Comment établit-on les faits linguistiques
- Qu'est-ce qu'un corpus ?

Qu'est-ce que la linguistique ? (1)

La linguistique moderne présentée dans ce cours s'est développée au début du XXème siècle, à partir des travaux de Ferdinand de Saussure (1857-1913).

Cette approche nouvelle pour analyser le langage humain n'est pas née de rien : de tous temps, le langage humain et les langues ont été étudiés. Si nous tentons de faire une histoire des théories du langage humain, nous pouvons distinguer plusieurs périodes avant l'apparition de la linguistique moderne :

Qu'est-ce que la linguistique ? (2)

- **L'Antiquité Grecque** : Dans la culture occidentale, la réflexion sur le langage est fortement marquée par la civilisation grecque classique. Plusieurs approches ou points de vue pour analyser le langage et les langues datent de cette période:
- La *rhétorique* où le langage est vu comme un moyen d'agir sur autrui;
- La *logique* (en particulier avec le langage) : cette réflexion philosophique tente d'articuler langage et vérité, de s'interroger sur les liens entre le monde réel et le langage (le rapport entre les objets du monde réel et les mots qui les désignent est-il arbitraire ou non, etc.) ;
- La *grammaire* : La première grammaire systématique de la culture occidentale est écrite par Denys de Thrace (-170 à -90). Il distingue les différentes parties du discours (Nom, article, adverbe, verbe, préposition, conjonction, etc.) et présente de façon systématique le fonctionnement de la langue grecque classique (déclinaison, conjugaison, etc.).

Qu'est-ce que la linguistique ? (3)

- **De l'Antiquité au XVIIème siècle** : Durant cette période, le Grec et le Latin sont considérés comme des modèles. Toute réflexion sur le langage et toute étude sur des langues particulières (même les langues vernaculaires : français, italien, etc.) se font selon les schémas hérités des grammairiens antiques. H. Etienne, par exemple, étudie en 1569 le français dans un ouvrage dont le titre est très éloquent, *Traité de la conformité du langage français avec le grec*.

Qu'est-ce que la linguistique ? (4)

- **Le XVIIème et XVIIIème siècle :** Durant cette période, les réflexions sur le langage et les langues se font dans une des perspectives suivantes :
- *La notion du "bel usage" ou "bon usage"* : Vaugelas publia en 1647 les *Remarques sur la langue française*, où il érige en norme les usages de la langue française faits à la Cour. L'étude de la langue a pour but d'imposer une norme.
- *La grammaire et les rapports entre langue et pensée* : L'ouvrage qui marque ce courant et cette période est la *Grammaire* dite de Port-Royal, écrite en 1660 par Arnauld et Lancelot. Dans cet ouvrage, l'étude des formes grammaticales se fait selon deux ordres :
- La description grammaticale du français : Dans une partie de l'ouvrage, les auteurs donnent une description du français.
- Le caractère universel du langage : Dans leur volonté de "logiciser" le langage, les auteurs essaient de montrer comment le fonctionnement du langage en général est en rapport étroit avec la logique de la pensée humaine.

Qu'est-ce que la linguistique ? (5)

Le XIXème siècle : Comparatisme et Linguistique historique : Durant cette période, la réflexion sur le langage est fortement influencée par la découverte du Sanskrit. Les chercheurs se consacrent principalement à l'évolution des langues dans le temps.

- *Le comparatisme* : Avec la découverte du Sanskrit, certains auteurs mènent des recherches dont le but est de montrer que des ressemblances importantes existent entre le Sanskrit et d'autres langues telles que le Latin, le Grec, le Persan, le Celtique, le Germanique, etc.
- *Romantisme et modèle biologique* : La langue devient un objet d'étude dans cette perspective "nationaliste" et historique: elle est considérée comme un organisme en évolution constante, marquée par l'histoire.
- *La linguistique historique* : Durant la seconde moitié du XIXème siècle, la phonétique connaît de grandes transformations et devient une science expérimentale. Ces progrès ont une influence importante sur la grammaire comparée: l'évolution historique d'une langue est analysée principalement dans une perspective phonétique. Cette nouvelle approche va permettre de définir des lois phonétiques fondamentales et d'avoir une analyse plus fine des changements phonétiques dans le temps.

Qu'est-ce que la linguistique ? (6)

La linguistique est une science récente qui a pour but d'étudier le langage humain, à partir de l'étude des multiples langues naturelles.

Cette définition étant donnée, la distinction entre *langue* et *langage* doit être faite :

- **le langage est la faculté humaine qui permet de communiquer.**
- **la langue** : la langue est ***la composante sociale du langage*** qui s'impose à l'individu. **Elle est un système de signes et de règles** reconnu par les membres de la communauté.

Qu'est-ce que la linguistique ? (7)

Pour « étudier les propriétés du langage humain à partir de l'étude des diverses langues », la linguistique s'est définie une méthode qui s'articule selon trois axes :

- **La linguistique est une science descriptive :** La description a pour but de comprendre le fonctionnement de la langue et du langage humain en général. Cela s'oppose donc à la grammaire traditionnelle que Saussure caractérise comme *normative*. Alors que le grammairien dicte des lois, le linguiste décrit et cherche à comprendre.

Qu'est-ce que la linguistique ? (8)

- **La linguistique reconnaît la primauté de l'oral sur l'écrit :** La linguistique ayant pour objet la langue vivante et parlée par une communauté, elle doit s'appuyer sur les données les plus immédiates, c'est à dire les données orales. Cette prise de position s'explique par le fait que:
  - la parole ou l'oral est premier;
  - les systèmes d'écriture sont une façon de "coder" la langue orale.

Qu'est-ce que la linguistique ? (9)

- **La linguistique privilégie l'approche synchronique :** La *synchronie* désigne un état de langue, la *diachronie* une évolution dans le temps. Dans une perspective synchronique, le fonctionnement de la langue est étudié à un moment donné, indépendamment de ce qui a pu se passer avant; en revanche, dans une perspective diachronique, la langue est étudiée dans son évolution en tenant compte de l'effet du temps sur elle. En accordant la primauté au point de vue synchronique, Saussure, et la linguistique structurale, ont montré une volonté de rompre avec la tradition linguistique

Qu'est-ce qu'un fait linguistique ? (1)

La description des faits linguistique passe par quelques mises au point ?

- L'opposition langue / parole ?
- Les niveaux d'analyse et de description de la linguistique.
- Les domaines de la linguistique

Qu'est-ce qu'un fait linguistique ? (2)

- Soit l'énoncé

Je viendrai demain

- Éléments sonores
  - Caractéristiques de la voix
  - Segments
  - Prosodie

Leur étude relève de la phonologie et de la phonétique, qui sont liés à l'opposition langue / parole.

Qu'est-ce qu'un fait linguistique ? (3)

Je viendrai demain

- Éléments lexicaux
  - je
  - Viendrai (venir)

LEXIQUE MORPHOLOGIE SYNTAXE SEMANTIQUE

Possibilité d'élargir vers la pragmatique.

Comment établir les faits en linguistique (1)

- L'étude des faits linguistiques passe par la collecte de données.
- Les données peuvent être collectées de différentes façons… chacune pouvant être déterminée en fonction de l'objectif visé.

Comment établir les faits en linguistique (2)

Deux méthodes distinctes pour décider quels sont les faits:

- *Utiliser un corpus :* On collecte un grand nombre d'énoncés produits par des locuteurs natifs du français; un **corpus.** On examine *ce qui a été dit pour en déduire ce qui peut se dire.*

- *Utiliser les jugements de grammaticalité :* Les locuteurs d'une langue ont une connaissance du système de leur langue qui leur permet de comprendre des énoncés et d'en produire d'autres. On utilise cette connaissance en demandant explicitement à des locuteurs de juger si certaines choses peuvent se dire ou non. C'est le **jugement de grammaticalité.** On examine *les jugements conscients des locuteurs pour* connaître *leur compétence inconsciente.*

Comment établir les faits en linguistique (3)

Il y a au moins quatre phénomènes qui rendent difficile la collecte des faits linguistiques:

- **La variation:** tout le monde ne parle pas tout le temps de la même façon
- **L'influence des variétés prestigieuses:** il y a des manières de parler/d'écrire qui sont mieux vues que d'autres
- **L'influence de la grammaire prescriptive**: on nous a appris qu'il y a des choses qu'il ne faut pas dire, même si les gens les disent en réalité.
- **Le culte du bon auteur**: l'idée que la littérature (de qualité) est la meilleure (voire la seule) source pour savoir ce qui est français ou pas.

Comment établir les faits ? (4)

- ! La variation

  Exemple en syntaxe : emploi des prépositions de lieu

  (1) a. *Je suis allé chez Marie.* [Paris]
  b. *Je suis allé à chez Marie.* [Auvergne]

  (2) a. *La mer monte jusqu'à la route.* [Paris]
  b. *La mer monte jusque la route.* [Bretagne]

Comment établir les faits ? (5)

- ! L'influence des variétés prestigieuses

(1) a. *Paul ne viendra pas.* b. *Paul viendra pas.*

(2) a. *Ne m'en donne pas.* b. *M'en donne pas.*

(3) a. *Donne m'en.* b. *Donnes-en moi*
c. *Donne m'en pas.* d. *Donnes-en moi pas.*
e. *Donne moi z'en pas.*

Comment établir les faits ? (6)

- L'influence des grammaires prescriptives :

Les grammaires traditionnelles fourmillent d'injonctions de la forme: « Ne dites pas X, dites Y ». Exemples :

(1) a. *Ne dites pas « aller au coiffeur », dites « aller chez le coiffeur ».*
(2) b. *Ne dites pas « aller en vélo », dites « aller à vélo ».*
(3) c. *Ne dites pas « après que je sois parti », dites « après que je suis parti ».*

Caractère commun de tous ces exemples: ils sont la marque de choses qui se disent effectivement.

L'interdiction coïncide parfois avec une variation socialement déterminée; parfois, elle ne correspond à rien de réel.

Comment établir les faits (7)

- De même qu'avec les corpus, les jugements ne font pas sans poser de problème :
  - Rien ne prouve que la connaissance inconsciente qu'ont les locuteurs de leur langue puisse être mobilisée consciemment. De fait, les individus non-entraînés ont bien du mal à produire des jugements de grammaticalité.
  - Soit on demande leurs jugements à des linguistes. Mais alors, comment être sûrs qu'ils ne sont pas influencés par leurs préconceptions théoriques?
  - Soit on entraîne des non-spécialistes. Mais comment être sûrs qu'on ne fausse pas leur jugement en les entraînant?

Comment établir les faits ? (8)

- On voit que la collecte des faits linguistiques ne va pas de soi. Le collectionneur est confronté à de nombreux phénomènes perturbateurs.

  NB: rien de ce qu'on a dit ne va à l'encontre de l'idée qu'il existe des faits linguistiques stables. Simplement, ceux-ci sont relatifs à une sous-communauté, à un temps, à une situation sociale, donnés.
- On est dans la situation normale dans les sciences empiriques: la collecte des données demande un appareillage et une méthode.
- Hypothèses pour avancer :
  - Le français est définissable comme un **ensemble d'énoncés possibles.**
  - Les locuteurs ont une **connaissance implicite de la langue** qui leur permet de la parler (plus ou moins) sans erreur.

Qu'est-ce qu'un corpus ? (1)

Le corpus est un ensemble homogène et significatif de données linguistiques observées et à partir desquelles pourra s'élaborer la description et la formalisation des faits linguistiques.

Il joue un rôle essentiel dans la linguistique structurale pour au moins une raison : dépasser la grammaire normative.

Qu'est-ce qu'un corpus ? (2)

Les caractéristiques d'un corpus significatif sont :

- L'homogénéité (le groupe qui le produit est socialement défini)
- La synchronie
- La moindre redondance possible

Avec le corpus, on reste en deçà de la perspective créative selon laquelle tout individu peut créer et comprendre des énoncés en nombre infini.

Exemples de corpus

- Travail sur les adjectifs à partir d'un relevé des adjectifs dans un dictionnaire.
- Recherche d'énoncés où les adjectifs sont placés différemment (avant le nom, après le nom, etc.) et ont diverses fonctions (épithète, attribut du sujet, attribut de l'objet, etc.)

Du corpus dans la linguistique à la linguistique de corpus

Avec la possibilité de stocker un nombre plus important de données, mais surtout de les traiter, la notion de corpus a évolué.

Nous allons tenter de voir en quoi consiste cette évolution, mais aussi quels sont les éléments essentiels à la construction d'un corpus.

Qu'est-ce qu'un corpus ? (3)

Définition généralement retenue :

La communauté linguistique considère, à la suite de Sinclair (1996), qu'un corpus est "*une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon de langage*".

D'après cette définition, un ensemble de données collectées ici et là sans réflexion préalable sur ce qui motive le rassemblement des documents n'est pas un corpus.

Qu'est-ce qu'un corpus ? (4)

Dans cette définition, plusieurs points font cependant débat :

1) que signifie *sélectionnées et organisées selon des critères linguistiques explicites* ? Un ensemble de données homogènes qui appartiennent à un *genre particulier* constitue-t-il un corpus (rassemblement de textes du *Monde*, rassemblement d'enregistrements d'émissions radiophoniques, etc.)

Qu'est-ce qu'un corpus ? (5)

2) Que signifie *pour servir d'échantillon de langage*?
Pour les tenants de la linguistique de corpus, les notions d'échantillonnage et de représentativité jouent un rôle essentiel, lorsqu'il n'est plus possible de rassembler de façon exhaustive toutes les formes répondant à l'objet d'étude (clôture du corpus). Mais, le passage de l'exhaustivité et de la clôture à l'échantillonnage et à la représentativité crée obligatoirement un décalage.

Qu'est-ce qu'un corpus (6)

----> toute collection de données même expérimentales peut être un corpus à part entière (choix des locuteurs, le nombre d'itérations et la sélection des formes peuvent être pensés de façon à gagner en représentativité).

Dans tous les cas, il faut avoir conscience des limites / des biais :
- le biais expérimental, - le biais du genre.

Qu'est-ce qu'un corpus ? (7)

Une autre définition plus souple a donc été proposée (Gibbon et al (1998) :

"A corpus is any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use of the data in-house, or by people in others organisations."