Informatique et Industrie de la langue
Les corpus

Pascal Amsili

Université Sorbonne Nouvelle

Février-Mars 2021

Ce diaporama reprend en partie des transparents utilisés par Kim Gerdès au cours des années scolaires précédentes.

Plan

Question de recherche

Comment fonctionne la faculté de langage ?

- Comment la décrire ?
- Comment la modéliser ?

Besoins

- cerveaux : pour élaborer des théories ;
- yeux : pour faire des observations

Besoins

- cerveaux : pour élaborer des théories ;
- yeux : pour faire des observations
- ... lectures : pour accéder aux connaissances accumulées

Chemins vers la connaissance linguistique

1. Introspection
   Risque : illusions, biais, manque de reproductibilité
2. Démontage du cerveau
   If you try and take a cat apart to see how it works, the first thing you have on your hands is a non-working cat. (Douglas Adams)

   Difficulté méthodologique : peut-on comprendre ce que vous écrivez dans votre traitement de texte en mesurant les variations d'énergie dans votre ordinateur ?
3. Observer les productions langagières
   - Expérimentalement (production induite, conditions contrôlées)
     Risque : conditions expérimentales non réalistes
   - "Naturellement" (recueil de productions, conditions écologiques)
     $\longrightarrow$ constitution de corpus
     Risque : données en vrac, des nombres et peu de compréhension

Sorbonne Nouvelle

Plan

Qu'est-ce qu'un corpus ?

Une collection de données linguistiques naturelles
textes écrits ou oraux

- Toute collection de textes est-elle un corpus ?

Qu'est-ce qu'un corpus ?

Une collection de données linguistiques naturelles
textes écrits ou oraux

- Toute collection de textes est-elle un corpus ? Non

Un corpus doit être :

- Représentatif
- Fini
- Numérisé
- Standard
- (Finalisé)

Définitions

- **Représentatif :**
  - **Ne pas utiliser un seul journal pour parler de la langue journalistique, ne pas utiliser un seul auteur pour parler de la littérature, ...**
- **Fini :**
  - **Corpus incrémentaux de plus en plus répandu. Mais : difficile à comparer les analyses effectuées sur différentes versions.**

Définitions

- **Informatisé**
  - **Avant l'informatisation existaient des corpus aussi, mais l'utilisation et l'extraction d'information étaient extrêmement laborieux**
- **Standard**
  - **Le corpus doit suivre les normes de la communauté :**
    - **Format, annotation, droits**
  - **Une collection de textes qui n'est pas utilisée par plusieurs personnes dans de recherches variées n'est pas vraiment un corpus. Une recherche sur un corpus qui n'est pas disponible à d'autres chercheurs n'est pas/difficilement vérifiable/falsifiable → elle est moins scientifique**

Définitions

- **Grand sujet :**
  - **Le partage de corpus**
  - **Qui travaille ? Qui paie ? Qui en profite ?**
  - **Ressource libre ?**
    - **Libre peut seulement vouloir dire “gratuit”**
    - **Mais normalement :**
      - **Libre = licence libre : Creative communs, …**

    → droit de modifier, exploiter, rediffuser avec attribution aux différents contributeurs.

Définitions

- **Autre grand sujet lié :**
  - **Protection de la vie privée**
    - **Comment faire un corpus**
      - **de SMS ?**
      - **De messages sur un réseau social ?**
      - **D'une conversation à table ?**
    - **Procédés d'anonymisation**

Corpus : oppositions pertinentes

- support : papier/numérisés/discrétisés
- écrit/oral/vidéo (lsf) ; multi-modaux
- monolingue/multilingue/aligné
- synchronique/diachronique
- statiques/dynamiques (incrémentaux)

- brut *vs.* annoté

  → Cours de L3 « Linguistique de corpus, annotation »

Plan

Linguistique et relation aux données

- Linguistique introspective (« de salon », "*armchair linguistics*")
  Les données sont "dans la tête du linguiste"
- Linguistique de corpus
  Les données sont "dans la nature"
- Linguistique expérimentale
  Les données sont fabriquées en laboratoire

Linguistique et relation aux données

- Linguistique introspective (« de salon », "*armchair linguistics*")
  Les données sont "dans la tête du linguiste"
- Linguistique de corpus
  Les données sont "dans la nature"
- Linguistique expérimentale
  Les données sont fabriquées en laboratoire

Attention aux caricatures !

Chemins vers la connaissance

- **Un vieux débat :**
  - **Le linguiste de corpus**

    vs.

  - **Le linguiste de fauteuil**
    - **Fillmore 1992**

Chemins vers la connaissance

- **Le linguiste de corpus**
  - **a tous les faits dont il a besoin, sous la forme d'environ un zillion de mots, et il conçoit son travail comme celui de dériver des faits secondaires à partir de ces faits primaires. En ce moment il est occupé à déterminer les fréquences relatives des onze parties du discours dans le premier mot d'une phrase et dans le deuxième mot d'une phrase**

    The corpus linguist has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus the second word of a sentence.

Chemins vers la connaissance

- **Le linguiste de fauteuil**
  - **est assis dans un fauteuil profond et confortable, les yeux fermés et les bras croisés derrière sa tête. De temps en temps, il ouvre un œil, se relève brutalement et crie : « Ouah ! Quel super fait ! », attrape son crayon, et écrit quelque chose. Il s'agite alors pendant plusieurs heures dans l'excitation de s'être encore approché d'une compréhension de la véritable nature du langage**

    The armchair linguist sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact!", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like.

Chemins vers la connaissance

- linguiste de corpus →linguiste de fauteuil
  - “Qu'est-ce qu'il pourrait me faire penser que ce que vous dites est vrai ?”
- linguiste de fauteuil →linguiste de corpus
  - “Qu'est-ce qu'il pourrait me faire penser que ce que vous dites est intéressant ?

Chemins vers la connaissance

- **Cas réel :**
  - **Chomsky:**
    - **The verb 'perform' cannot be used with mass word objects: one can 'perform a task' but one cannot 'perform labour'**
  - **Hatcher:**
    - **How do you know, if you don't use a corpus and have not studied the verb 'perform'?**
  - **Chomsky:**
    - **How do I know'? Because I am a native speaker of the English language**
- **Or, d'après le corpus BNC, on peut 'perform magic'**

Liguistique introspective

- tire profit de l'intuition du chercheur (linguiste)
- est immédiatement accessible
- offre des exemples négatifs (contrefactuels)
- permet une étude systématique des variations → linguistique expérimentale
- permet de s'éloigner des influences extra-linguistiques → linguistique expérimentale
- tire profit de l'attention du linguiste → linguistique de corpus

mais...

- est potentiellement sensible au biais de l'expérimentateur
- n'est pas protégée contre l'influence du dialecte/idiolecte

Linguistique de corpus

- oriente vers une linguistique de l'usage
- offre toute la puissance des outils statistiques
- permet d'objectiver les observations
- permet de répliquer les observations/manipulations
- permet de faire une démarche "bottom-up"

mais...

- n'offre pas de données négatives :

  l'absence de preuve n'est pas la preuve de l'absence

- dépend de la capacité de collecter des données en grand nombre

Linguistique expérimentale

- permet de contrôler les variables ayant une influence
- permet de "purifier" les phénomènes observés
- permet la réplication et la cumulativité en science
- permet d'objectiver les observations

mais...

- coût très élevé : chaque variation de paramètres nécessite de nouvelles expérience
- suppose une théorie faisant des prédictions ("top-down")
- demande qu'on vérifie le passage du laboratoire au contexte écologique

Plan

Trois périodes

- Avant Chomsky $\rightarrow$ 1955
- Domination de la vision chomskyenne 1955 $\rightarrow$ 1985
- Emergence d'une nouvelle linguistique de corpus

Fréquences

Qu'est-ce qu'on observe quand on regarde un corpus ?

- Avant tout des **fréquences** !
  - Lord Kelvin : « mesurer c'est savoir »

Fréquences

Qu'est-ce qu'on observe quand on regarde un corpus ?

- Avant tout des **fréquences** !
  - Lord Kelvin : « mesurer c'est savoir »
- Quelle phrase est la plus fréquente ?

Il habite à Trifouilly-les-Oies	
Il habite à Shanghai	
Il habite à Paris	

Fréquences

Qu'est-ce qu'on observe quand on regarde un corpus ?

- Avant tout des **fréquences** !
  - Lord Kelvin : « mesurer c'est savoir »
- Quelle phrase est la plus fréquente ?

Il habite à Trifouilly-les-Oies	1
Il habite à Shanghai	1 860
Il habite à Paris	66 300

Fréquences

Qu'est-ce qu'on observe quand on regarde un corpus ?

- Avant tout des **fréquences** !
  - Lord Kelvin : « mesurer c'est savoir »
- Quelle phrase est la plus fréquente ?

Il habite à Trifouilly-les-Oies	1
Il habite à Shanghai	1 860
Il habite à Paris	66 300

Et ces trois phrases sont grammaticales

Fréquences

Qu'est-ce qu'on observe quand on regarde un corpus ?

- Avant tout des **fréquences** !
  - Lord Kelvin : « mesurer c'est savoir »
- Quelle phrase est la plus fréquente ?

Il habite à Trifouilly-les-Oies	1
Il habite à Shanghai	1 860
Il habite à Paris	66 300

Et ces trois phrases sont grammaticales

- Chomsky : Any natural corpus will be skewed. Some sentences won't occur because they are obvious, other because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list (1962)

Histoire

- **En un mot : Chomsky dit qu'un corpus est toujours**
  - **Partial**
  - **Partiel**
- **Dans l'opposition**
  - **Performance / compétence**

  il ne veut décrire que la compétence

Performance / Compétence

- L'opposition date du structuralisme :
  - Ergon – Energeia (Wilhelm von Humboldt)
  - Sprache – Rede (Hermann Paul 1880)
  - Sprachsystem – aktualisierte Rede (Georg von der Gabelentz 1891)
  - Langue – Parole (Ferdinand de Saussure 1896)
  - Sprachgebilde – Sprechakt (Karl Bühler 1934)
  - register – use (Michael A. K. Halliday 1961)
  - Compétence – performance (Noam Chomsky 1965)
- La linguistique de corpus essaie depuis les années 1980 de dépasser cette opposition en s'intéressant à
  - l'usage

Histoire

- **1900 – 1950 :**
  - **Structuralisme**
  - **Linguistique descriptive**
  - **Behaviourisme**
  - **Étude de relations entre les objets d'étude, intégrée dans une étude de tout**
  - **Courant intellectuel qui prônait la neutralité par rapport à l'objet de l'étude**
  - **Nécessité de sauver le mémoire des langues amér-indiennes**

→ impossibilité/interdiction d'introspection

Histoire

- **Après les critiques de Chomsky, les corpus survivent surtout dans des domaines de la linguistique où l'introspection est impossible :**
  - **L'acquisition**
  - **L'étude des langues anciennes**

  Et à un moindre mesure :
  - **La sociolinguistique (mesure de variations, …)**
  - **La phonologie**

Histoire

- Le retour du corpus dans les années 1980
  - Existence d'outils : L'ordinateur personnel
  - Hors monde linguistique :
    - Littéraires
    - Ingénieurs en TAL :
      - “Every time I fire a linguist, the performance of the speech recognizer goes up” (Frederick Jelinek, ~1988)