Informatique et Industrie de la langue
Les corpus

Pascal Amsili

Université Sorbonne Nouvelle

Janvier 2020

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

Plan

Data-driven linguistics

Linguistique de corpus

Histoire de la linguistique de corpus

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)