Pas d'outils sans approche sous-jacente : Le concordancier et le contextualisme

Geoffrey Williams

MCF (Militant des Corpus en France)

Université de Bretagne Sud

Je ne suis pas, Je suis

- Pas un linguiste
  - Le formalisme est réducteur
- Peut être un linguiste appliqué
  - L'environnement social est crucial
- Linguiste de corpus
  - Un explorateur de la communication linguistique
  - Pas de jugements, pas de compromis, j'observe
- Pas politiquement correct

Un euro - cent centimes

- L'article 2 du règlement européen du 3 mai 1998 dispose : "Un euro est divisé en cent cents". Ce même règlement précise que le nom "cent" n'empêche pas l'utilisation de variantes de cette appellation dans la vie courante dans les Etats membres".

  La revue fiduciaire. 3 février 2001 : 21.

Sauver la langue

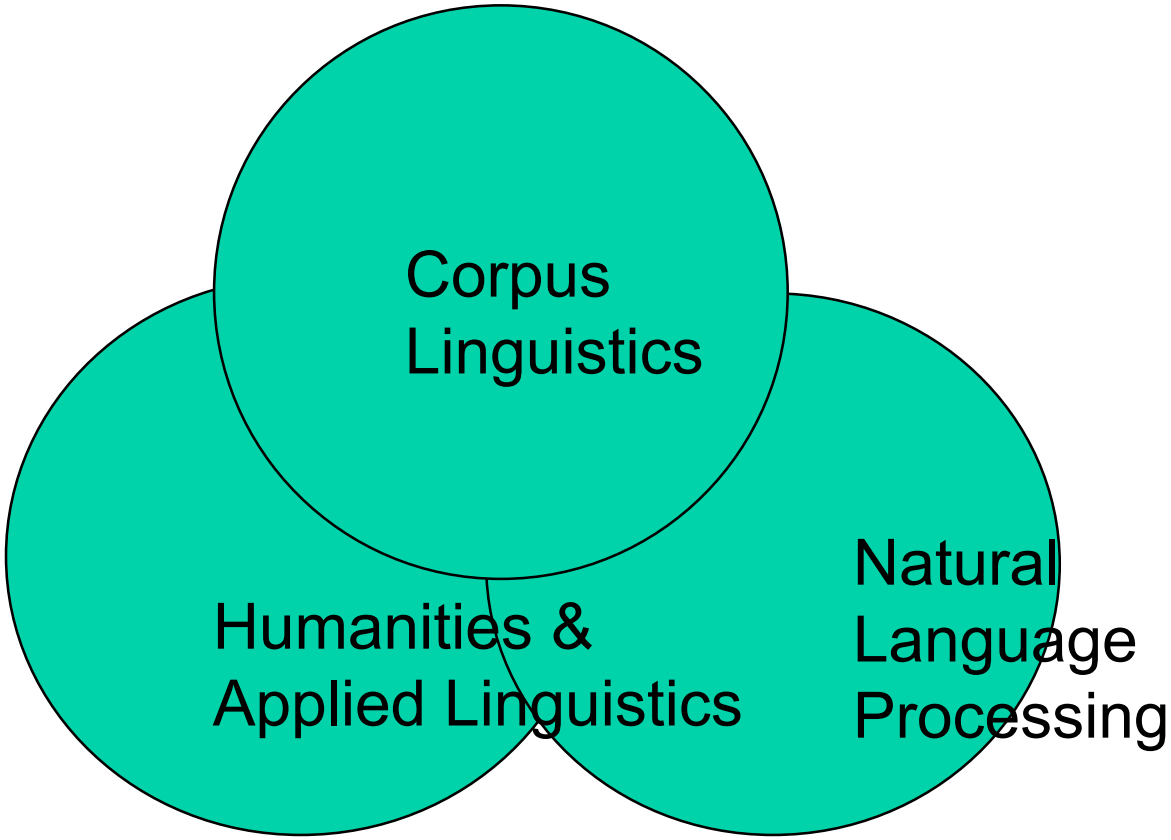
- Sauver la langue
  - Quelle langue, par qui et pourquoi?
- Pivot et les mots en perdition
  - A quoi bon?
- Les diktats parisien – l'académie des morts
  - La langue = la littérature
  - Le français correct = Le Monde

Corpus-driven linguistic analysis

- We must remember :….
- (b) That the Latin peoples always think *a priori* and by deduction. This is the tradition of Aristotle and St Thomas Aquinas. Since the Reformation Anglo-Saxons think *a posteriori* and inductively

(MacMillan, H, 1969, Tides of Fortune, London : MacMillan Press. P193.)

Common Ground



Une typologie

- La linguistique de corpus est tournée vers l'extérieur
- La lexicométrie est tournée vers l'intérieur
- Les deux s'intéressent à la parole
- Le NLP est tourné vers les outils
- Les outils peuvent être les mêmes, l'approche est différente

La linguistique de corpus

- Qu'est-ce que c'est?
  - Qu'est ce qu'un corpus
  - Les origines
  - Les approches
- A quoi ça sert?
  - L'enseignement, la lexicographie, la terminologie
- Les bases
  - Des concordanciers – WordSmith et Wconcord
  - SARA – pour aller plus loin

Contexte

- Contexte de culture
- Contexte de situation
- *without some imperative stimulus of the moment, there can be no spoken statement. In each case, therefore, utterance and situation are bound up inextricably with each other and the context of situation is indispensable for the understanding of the words* (op. cit. 307)."
- MALINOWSKI B. (1924) The Problem of Meaning in Primitive Languages in OGDEN & RICHARDS (1924) p296-336.

L'héritage de Firth

- You shall know a word by the company it keeps. (Firth 1957)
- The basic assumption of the theory of analysis by levels is that any text can be regarded as a constituent of a *context of situation.*
- Firth. 1957. A synopsis of linguistic theory 1930-1955.

Firth et l'école de Londres

- Eugène Winter, spécialiste d'analyse de discours,
- Michael Hoey – linguiste de corpus, élève de Winter
- Michael Halliday, le grammairien,
- John Sinclair, grand spécialiste de la linguistique de corpus et fondateur du projet COBUILD

L'approche contextualiste (1)

- La contribution de la lexicographie
  - *Dictionnaire de l'Académie Française* 1694
    - normaliser et fixer le français
  - Samuel Johnson 1755
    - décrire l'anglais de son époque

L'approche contextualiste (2)

- La linguistique
  - Saussure, Bloomfield, Chomsky
    - une linguistique formelle
  - Malinowski, Wittgenstein, Firth
    - contexte de culture, contexte de situation
    - le sens contextuel, les prototypes
    - une théorie distributionelle

l'informatique et les sciences humaines

- Le contextualisme britannique
  - l'école de Londres
  - l'école de Birmingham
- Data-driven research - un état d'esprit
- Les sciences humaines aux Etats-Unis

Les Origines

- Brown Corpus
- Lancaster-Oslo-Bergen (LOB) Corpus
- London-Lund
- COBUILD au British National Corpus (BNC)

Ce que c'est

- La linguistique de corpus et les corpus
- La linguistique de corpus et les ordinateurs
- La linguistique de corpus et la langue
- La linguistique de corpus et les régularités
- La linguistique de corpus et la quantité
- La linguistique de corpus et la comparaison

La linguistique de corpus et les corpus

- Qu'est qu'un corpus?
  - Des grands ensembles de textes selectionés
  - Des textes entiers ou des échantillons importants
- Un corpus n'est pas
  - Des phrases isolées
  - L'œuvre d'un auteur
  - Une collection hétérogène de textes
  - Des pages web
  - Trois ou quatre articles sur un thème

La linguistique de corpus et les ordinateurs

- Les corpus sont électroniques
- Les ordinateurs sont des outils, nous ne sommes pas leur esclaves
- Des outils prêts à tourner
  - Wconcord
  - WordSmith
  - SARA - XARA
- La Programmation
  - Essayer PERL, si vous avez vraiment envie.

La linguistique de corpus et la langue

- Des textes authentiques en contexte
  - En entier ou en échantillon
- La langue et la communication
  - Comment des gens communiquent
  - Les mots et la syntaxe sont liés
- Le but de la langue, la communication
  - Ce n'est pas pour tester des outils
  - Les linguistes de corpus aiment la langue

La linguistique de corpus et les régularités

- La linguistique de corpus et la lexicométrie
  - La linguistique de corpus est tournée vers l'extérieur
  - La lexicométrie est tournée vers l'intérieur
- Des régularités et les statistiques
  - Pouvoir généraliser
- On se répète
  - Les idiomes
  - Les collocations
  - Les phrases lexicales
  - Les grammaires locaux

La linguistique de corpus et la quantité

- Les études pilotes et les petits corpus
- Les ressources textuelles
- Homogénéité
- Les corpus: la taille, la variété et la choix
  - Petit - entre 500,000 – 1 million d'occurences
  - 150+ textes
  - choice, not chance (Engwall, 1994)

La linguistique de corpus et la comparaison

- Comparaison entre des corpus de référence et des corpus spécialisés
- Comparaison entre des genres
- Comparaison entre des listes
- Comparaison entre des langues

Construire soi-même

- Soyez nevrosé
- Est ce que c'est assez grand?
- Quels critères de sélection?
  - Critère externes – discipline ou communauté de discours
  - Critère interne – la représentativité
- Où trouver des textes
  - internet, numériser des textes, taper à la machine
- Les droits d'auteur

Les approches

- Corpus-based studies
  - Une approche empirique et déductive
- Corpus-driven studies
  - L'approche inductive
  - *Linguistic bungee jumping*

Les concordanciers

- Indexation

  l'évolution de la vitesse

  Wconcord – WordSmith

  pas d'indexation – opérations « simples »

  SARA

  indexation – fichier lourds avec analyse morphosyntaxique et balisage XML

Les opérations de base

- Fichiers traités
  - Textes balisés – html ou XML
  - Analyse morphosyntaxique
- Listes de fréquence
- Concordances
- Recherche simple ou complexe
- Expressions régulières
- Les collocations – co-occurrences ou calculs statistiques

Un peu plus complexe

- Gestion des textes avec des métadonnées
- Comparaison de fichiers ou de listes
- Textes en parallèle ou alignés
  - Paraconc

Moins bruts, plus sophistiqués

- Text Encoding Initiative
  - Balisé des textes
    - Garder les caractéristiques des documents,
    - stocker l'information sur des sources
  - SGML
    - Complet, complexe, manque de navigateurs
  - XML
    - Aussi complet, moins complexes, navigateurs web standards
- Analyse morphosyntaxique
- Lemmatisation