Traitements de corpus

John Carroll, Laura Kallmeyer

laura.kallmeyer@linguist.jussieu.fr

1er semestre 2000–2001

Page web du cours: http://talana.linguist.jussieu.fr/∼lkallmey/LI031.html

1 Introduction

Les linguistiques de corpus: une méthodologie plutôt qu'une branche de linguistique comme la syntaxe, la sémantique etc.

1.1 Histoire des linguistiques des corpus

(McEnery and Wilson 1996 chapitre 1)

Avant Chomsky, la plupart des recherches linguistiques étaient basées sur des corpus.

Les critiques principales de Chomsky (1957, 1965) aux approches basées sur des corpus:

1. Le but de la recherche linguistique est de décrire notre compétence linguistique plutôt que la performance. Or il faut décrire notre propre connaissance interne d'une langue. La performance est l'évidence externe de la compétence et elle est influencée par des facteurs en dehors de la compétence (comme fatigue, mauvaise concentration, ivresse etc.).

   Un corpus reflète la performance et par conséquent ne peut jamais donner un image fiable de la compétence.

2. Le nombre des phrases d'un langage naturel est infini: *J'ai vu hier l'homme qui a écrit le livre dont j'ai lu seulement la fin qui parlait d'une jeune fille qui...*

   Pour cette raison on ne peut jamais collectionner toutes les phrases d'un langage dans un corpus. Il y a toujours des phrases qui ne s'y trouvent pas mais qui sont parfaitement grammaticales.

3. L'introspection peut donner des résultats qu'on ne peut pas obtenir par un corpus:
   - Elle peut donner des jugements sur n'import quelle phrase, mêmes les données qui ne figurent pas dans un corpus.
   - Elle peut donner des jugements négatifs. Par contre, le fait qu'une phrase ne figure pas dans un corpus ne signifie pas qu'elle n'est pas grammaticale.

Une autre critique était que des recherches dans des corpus prenaient beaucoup de temps et étaient très coûteux avant l'accessibilité des corpus éléctroniques. (Cela ne pose plus un problème aujourd'hui.)

Les critiques de Chomsky ne sont qu'une partie de la vérité.

Des problèmes avec des données introspectives:

1. Des données de corpus peuvent être facilement observées et vérifiées tandis que la vérification des données introspectives et beaucoup plus difficile.

2. Des données introspectives sont des données artificielles.

3. Par introspection on ne peut pas obtenir des résultats sur la fréquence de certains mots ou de certaines constructions.
4. Dans quelques domaines de linguistique comme l'acquisition des langues on ne peut pas obtenir des jugements introspectives, par conséquent on y doit utiliser des corpus.
5. Le même est le cas pour les langues mortes. L'absence des jugements oblige le linguiste à se servir des corpus.

Tous les deux, l'approche introspectif et l'approche basé sur des corpus, ont des avantages. L'idéal est de se servir de tous les deux.

Après la critique de Chomsky jusqu'aux années 80, seulement très peu de recherches linguistiques suivaient un approche de corpus. Les plus importants corpus de cette époque sont le SEU et le Brown Corpus (cf. 1.4).

1.2 Qu'est-ce que c'est qu'un corpus?

(McEnery and Wilson 1996 chapitre 2)

Un corpus est une collection de textes qui sert d'échantillon d'un langage. En général, les textes d'un corpus sont sélectionnés et organisés selon des critères linguistiques explicites.

Aujourd'hui on entend par "corpus" toujours une collection de textes éléctronique.

Dans le cas idéal, un corpus a les qualités suivantes:

1. **Représentativité**: Les textes d'un corpus doivent être choisis d'une telle façon qu'ils constituent un échantillon représentatif du langage en question.
2. **Taille finie**: En général, un corpus est une collection finie de textes (1 million mots par exemple).

   Exceptions: corpus moniteurs (monitor corpus).
3. **Textes éléctroniques**: Un corpus est en général une collection de textes éléctronique qui peut être manipulée (annotée, utilisée pour des recherches) par ordinateur.

1.3 Annotation des corpus

(McEnery and Wilson 1996, chapitre 2, Garside et al. 1997, chapitre 1)

Il faut ajouter des charactérisations abstraites aux éléments d'un corpus pour pouvoir l'utiliser d'une façon efficace pour des rechercches linguistiques. Ces charactérisations supplémentaires s'appellent des annotations.

Exemple 1 *Suppose qu'on cherche des données qui manifestent l'utilisation de "comme" comme comparative 'réduite' (comparatif sans verbe).*

On veut trouver des phrases comme

(a) Elle chante comme un rossignol
(b) Comme violoniste on ne trouve pas mieux

Mais quand on cherche des occurrences de "comme" dans un corpus, on va aussi trouver des utilisations de "comme" des types suivants:

(a) Comme il faisait beau, je suis sortie
(b) Jean comme Pierre sont malades

Dans un corpus étiqueté avec des partie du discours, on peut distinguer les trois emplois de comme. Par exemple dans le corpus Le Monde (TALaNa), on a les annotations suivantes:

(a) Elle chante comme:P un rossignol (P = préposition)
(b) Comme:CS il faisait beau, je suis sortie (CS = conjonction de subordination)
(c) Jean comme:CC Pierre sont malades (CC = conjonction de coordination)

Exemple 2 *Suppose qu'on veut trouver dans un corpus allemand des constructions où un syntagme prépositionnel topicalisé modifie un objet direct comme dans*

über Chomsky habe ich ein Buch gelesen
sur Chomsky ai je un livre lu
'J'ai lu un livre sur Chomsky'

Pour chercher des telles constructions, il faut avoir accès aux structures syntaxiques, particulièrement aux relations de dépendance: on doit chercher des phrases qui commençent par un syntagme prépositionnel et qui contiennent un objet direct dont le syntagme prépositionnel est un modificatuer.

Pour cela il faut une annotation syntaxique.

Les avantages d'une annotation linguistique:

1. Extraire de l'information (voir exemples)

   Pour rechercher dans un corpus annoté, on peut

   (a) souvent plus facilement formuler des requêtes adéquates, et

   (b) formuler des requêtes ciblées qu'on ne peut pas formuler sans l'annotation.

2. Ré-utilisabilité

   Un corpus annoté peut être utilisé pas seulement par celui (ou ce groupe) qui a exercé l'annotation mais aussi par d'autres chercheurs.

3. Multi-fonctionalité

   L'annotation d'un corpus peut servir à des fins complètement différentes, souvent pas prévisibles au moment de l'annotation.

Les différents niveaux d'annotation:

- Informations textuels et extra-textuels

  Informations sur le texte ou le corpus:

  1. le titre, l'auteur et la date de la publication d'un texte
  2. le langage d'un texte
  3. la catégorie d'un texte (fiction – non-fiction etc.)
  4. la taille d'un corpus (en mots, kByte etc.)

     ...

  Ces informations sont normalement encodées dans le *header* d'un corpus.

- Orthographie

  Si possible, utiliser seulement l'alphabète ISO-646 pour guarantir la ré-utilisabilité d'un corpus.

  Recommendations pour l'encodage des accents par *Text Encoding Initiative (TEI)*. Par exemple

  `é` pour "é", `â` pour "â", `&uumlaut;` pour "ü" etc.

- Segmentation (*tokenization*)

  Il s'agit de segmenter un corpus en unités ("mots") qui ne correspondent pas toujours à des mots graphiques.

  > des cours ⇒ de + les + cours (des est une préposition plus article défini) ou de_les cours (des est un déterminant composé plus article défini)
  > en dépit de ⇒ un "mot" (préposition)

- Etiquetage avec parties du discours (*part-of-speech tagging*)

  Attribuer des catégories morpho-syntaxiques (parties du discours) aux mots d'un corpus.

  > Il:CL était:VI tel:A que:CS tu:CL l':CL avais:VI laissé:VK

- Lemmatisation

  Enrichir un corpus avec des lemmes (par exemple "venir" pour *viens, venons, venant, ...*).

- Annotation syntaxique (*parsing*)

  Annoter le corpus avec des structure syntaxique, par exemple (style de Penn Treebank)

```
(S (NP-SBJ-1 Jean)
   (VP doit
       (S (NP-SBJ *-1)
          (VP avoir
              (VP rendu
                  (NP le livre))))))
```

- Annotation sémantique

  Désambiguïsation lexicale: Annoter les mots d'un corpus avec des sens ou avec des champs sémantiques.

1.4 Exemples des corpus

(Kennedy (1998), section 2.3)

Corpus anglais:

SEU "Survey of English Usage", le corpus pré-éléctronique le plus important. Commencé par Quirk en 1959. Anglais britannique écrit et oral. 100 textes écrits, 100 textes oraux, ca. 5000 mots par texte. Etiqueté. Aujourd'hui le corpus existe en forme éléctronique.

Brown Corpus commencé en 1961 par Francis et Kucera à Brown University. Corpus d'anglais américain écrit d'un million de mots, premier corpus éléctronique. Etiqueté. Disponible en 1964.

BNC corpus représentatif de l'anglais britannique écrit et oral, compilé entre 1991 et 1996. 100 millions mots, 4124 textes dont 90% écrits et 10% oraux. Automatiquement annotés avec des parties de discours (*part of speech*).

Penn Treebank (University of Pennsylvania). Corpus arboré de 4 millions de mots d'anglais américain. Contient le Brown Corpus.

Corpus français:

Frantext Textes français, Institut National de la Langue Française. Environ 3500 oeuvres du XVIème au XXème siècle, 80% d'oeuvres littéraires, 20% d'oeuvres techniques. Une partie du corpus est étiqueté.

Le Monde – TALaNa Corpus Le Monde, étiqueté à TALaNa, 1 million mots environ, annotation syntaxique en cours

1.5 Standards d'annotation

Encodage avec SGML ou XML.

```
<G>Au_cours_de</G><L>au_cours_de</L><EE>P</EE>
<G>la</G><L>le</L><EE>Ddef-fs</EE>
<G>conf&eacute;rence</G><L>conf&eacute;rence</L><EE>NC-fs</EE>
...
```

Des standards d'annotation servent à guarantir

1. qu'on peut utiliser des outils standard pour annotation et recherche du corpus,
2. la ré-utilisabilité du corpus, et
3. la comparabilité avec d'autres corpus.

Des standards pour l'encodage des corpus avec SGML étaient développés par le TEI (Text Encoding Initiative) et par EAGLES (Expert Advisory Group on Language Engineering Standards).

1.6 Choix terminologiques

(Habert et al. 1997, Introduction)

La *précision* (*precision*) représente la proportion de réponses pertinentes extraites par rapport au total des réponses extraites, le *bruit* est la proportion de réponses non-pertinentes extraites par rapport au total des réponses extraites.

Le *rappel* (*recall*) est la proportion des réponses pertinentes extraites par rapport auf total des réponses pertinentes possibles, le *silence* est la proportion des réponses pertinentes non-extraites par rapport au total des réponses pertinentes possible.

Exemple 3 *Suppose qu'on veut extraire des occurrences de "comme" comme comparatif réduit d'un corpus contenant 440 occurrences de "comme" dont 240 comme comparatif réduit.*

Si la requête extrait toutes les occurrences de "comme", on a les valeurs suivantes:

Précision: $\frac{240}{440} = 55\%$

Bruit: $\frac{200}{440} = 45\%$

Rappel: $\frac{240}{240} = 100\%$

Silence: $\frac{0}{240} = 0\%$