Introduction à l'utilisation des corpus

Alexandra MESTIVIER
alexa.mestivier@gmail.com

1. Qu'est-ce qu'un corpus?

Question ouverte

Est-il bien utile d'utiliser les corpus comme aide à la traduction étant donné l'investissement/l'effort nécessaire pour les construire ?

Questions abordées

Qu'est-ce qu'un corpus ?

- Quels types de corpus y a-t-il ?
- Quelques exemples.
- A quoi peut servir un corpus ?
- Dans quel but doit-on constituer des corpus dans le cadre du Master ?
- Comment stocker le corpus ?

Questions abordées

- **Les prochaines séances**
  - Sous quelle forme faut-il le stocker pour qu'il soit facilement utilisable ? (rappel sur les formats de fichiers)
- De quels outils dispose-t-on pour exploiter les corpus?

  Outils d'interrogation des corpus

Qu’est-ce qu’un corpus?

Les corpus sont des

- collections de textes **de taille importante**

  (BNC=100 Million words !)

- constituées de textes **authentiques**
- rassemblées selon des **critères spécifiques**
- collectées **sous format électronique**.

Corpus et concordanciers

- **Format électronique ➔**
  les corpus ne sont pas faits pour être consultés de manière séquentielle (~livre) mais interrogés (concordanciers)

- **Un concordancier est un logiciel qui construit des concordances.**

- **La plupart des logiciels d'analyse textuelle sont bases sur le format texte brut (.txt).**
  Pas de

ET A QUOI RESSEMBLE UNE CONCORDANCE ?

Concordances monolingues

Exemple 1 Exemple 2

- **comparer les divers emplois|sens d'un même terme**
- **observer la fréquence des mots**
- **identifier des collocation, définitions**
- **observer des propriétés distributionnelles de certains mots.**

•Outils dérivés : les Voisins de le Monde, Word Sketch

Concordances bi-lingues

Exemple 1

Exemple 2

- la traduction des passages correspondant à la requête
- résoudre les problèmes de traduction que d'autres traducteurs ont déjà rencontrés???
- méthodes d'alignement

- Mémoires de traduction
- Entrée aux systèmes de traduction automatique

Quels types de corpus existe-t-il?

- **support** : papier, électronique, oral, vidéo
- **version langagière** :
  - monolingue, bilingue (comparable ou alignés), multilingue
  - originaux, traductions
  - locuteurs natifs ou apprenants de la langue
- **état de la langue** : synchronique ou diachronique
- **but** : corpus de référence ou de spécialité.
- **ouvert // ferme**
- **présence d'** **annotés**

QUEL TYPE D'ANNOTATION????

Quel type d'annotation ?

- les attributs de formatage : paragraphes, sections, titres, etc.
- l'information textuelle : date de publication, auteur, type de texte, registre, etc. Exemple
- l'analyse linguistique du contenu du texte :
  - étiquetage morpho-syntaxique (tagging) Exemple
  - lemmatisation Exemple
  - analyse syntaxique Exemple

Dans quels domaines on les utilise?

- Lexicographie (aide a la constitution de dictionnaires)
- Apprentissage des langues
- Études sociolinguistiques
- Linguistique : (l'étude de vocabulaire, de la grammaire, évolution de la langue ou des sens des mots.
- Linguistique informatique (TALN), entraîner ou tester les outils d'analyse textuelle
- Terminologie, traduction, rédaction technique
  - analyser les caractéristiques des textes traduits.
  - aide à la traduction.

Réflexion

Quels sont les avantages des corpus par rapport aux

- **Textes imprimes**
- **Dictionnaire (hint)**
- **Expert**
- **WWW**
- **Intuition**

- Exemple : mot *umbrella* dans Oxford English Dictionary :

1/ portable protection against rain, consisting of a circular piece of fabric mounted on a foldable frame of spikes attached to a central stick that serves as a handle.

2/ Any kind of general protecting force or influence.

Comparer avec l'information dans le BNC