Corpus et méthode expérimentale

module X9IT080

Florian Boudin

Département informatique, Université de Nantes

Révision 1 du 30 juillet 2012

Présentation du module

- Notions abordées dans ce module
  - ...
- Volume horaire : 15 séances (20h)
  - 6 CM (8h) - 6 TD (8h) - 3 TP (4h)
- Notation

Plan

Définitions I

Corpus

Un **corpus** est un ensemble de documents (textes, images, vidéos, etc.) regroupés dans une optique précise.

Dans le cadre de ce cours : corpus → corpus de textes

Plusieurs caractéristiques sont à prendre en compte pour la création d'un corpus bien formé :

- la taille ;
- le langage ;
- le temps couvert par les textes du corpus ;
- le registre de langage.

Source : `http://fr.wikipedia.org/wiki/Corpus`

Définitions II

Taille

Le corpus doit évidemment atteindre une taille critique pour permettre des traitements statistiques fiables.

- Si le corpus est utilisé pour construire des modèles de langue, quelle doit être sa taille minimum ? 1M mots, 10M mots, etc.

Langage

Un corpus **monolingue** bien formé doit nécessairement couvrir une seule langue, et une seule déclinaison de cette langue (e.g. français de France et français du Québec).

Définitions III

Période couverte

Le temps joue un rôle important dans l'évolution du langage : le français parlé aujourd'hui ne ressemble pas au français parlé il y a 200 ans ni, de façon plus subtile, au français parlé il y a 10 ans, à cause notamment des néologismes.

Registre de langage

Un corpus construit à partir de textes scientifiques ne peut être utilisé pour extraire des informations sur les textes vulgarisés, et un corpus mélangeant des textes scientifiques et vulgarisés ne permettra pas de tirer de conclusion sur ces deux registres.

Corpus parallèle

Un corpus parallèle est un ensemble de paires de textes tel que, pour une paire, un des textes est la traduction de l'autre.

Construire un corpus parallèle nécessite un **alignement** des unités textuelles :

- mettre en correspondance des unités textuelles en langue source avec celles de la langue cible.

L'alignement des unités textuelles peut être manuel ou automatique.

La granularité de l'alignement (documents, phrases, mots) dépend de l'utilisation du corpus :

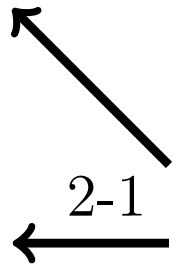
- traduction automatique, génération de paraphrases, construction dictionnaires bilingues, etc.

Alignement de phrases

Il s'agit de l'un des sauropodes les plus connus.

C'était un très grand quadrupède au long cou, avec une longue queue en forme de fouet.

Ses membres antérieurs étaient légèrement plus courts que ses membres postérieurs, ce qui lui donnait une posture horizontale.



2-1

1-1

0-1

One of the best-known sauropods, Diplodocus was a very large long-necked quadrupedal animal, with a long, whip-like tail.

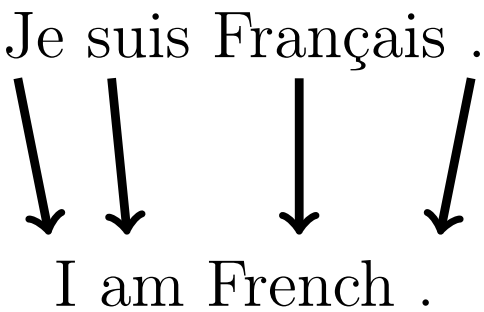
Its forelimbs were slightly shorter than its hind limbs, resulting in a largely horizontal posture.

It is the longest dinosaur known from a complete skeleton.

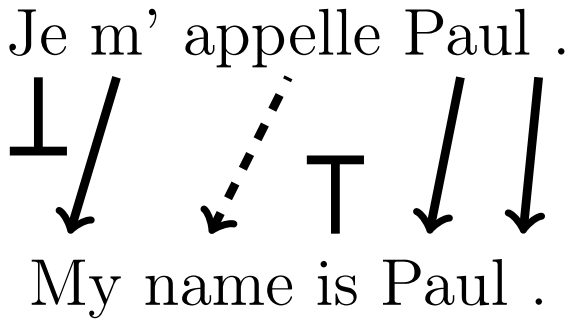
Source : `http://en.wikipedia.org/wiki/Diplodocus`

Alignement de mots

Exemple d'alignment simple.



Exemple d'alignment plus compliqué.



Corpora parallèles disponibles

- Europarl
  Délibérations du Parlement européen disponibles en 21 langues.
- OpenSubtitles
  Sous-titres de films disponibles en 30 langues.
- Hansard (Canada)
  Rranscriptions officielles des débats parlementaires dans les gouvernements.

References I