cqpweb.univ-avignon.fr

Présentation de l'interface d'interrogation de corpus 'cqpweb'. Graham Ranger.

Séminaire Agor@ntic, 7 décembre 2016.

cqpweb.univ-avignon.fr

- Introduction
- Historique
- Architecture
- Utilisation… et utilité
- … et l'interdisplinaire

cqpweb.univ-avignon.fr

cqpweb : introduction

- cqp : corpus query processor
- web : en ligne

Interface web pour l'interrogation de grands corpus de textes.

Ce semestre : projet pédagogique de préparation de corpus, par des étudiants de M1 et M2.

cqpweb.univ-avignon.fr

cqpweb : l’historique

Au départ : British National Corpus

≈ 100.000.000 mots d’anglais britannique

L’un des premiers megacorpus, constitué dans les années 1990

Plus de 4000 textes différents, classés par type, genre, source, etc.

cqpweb.univ-avignon.fr

cqpweb : l’historique

British National Corpus :

D’abord un corpus privé, payant ;

Puis interrogeable via différentes interfaces ;

Puis gratuit et entièrement téléchargeable.

cqpweb.univ-avignon.fr

cqpweb : l'historique

L'une des interfaces du BNC s'appelle BNCweb : http://bncweb.lancs.ac.uk/ (Hoffmann et al 2008.)

Interface d'interrogation écrite en perl.

Avantages : souplesse et puissance : utilisation de cql corpus query langage, notamment.

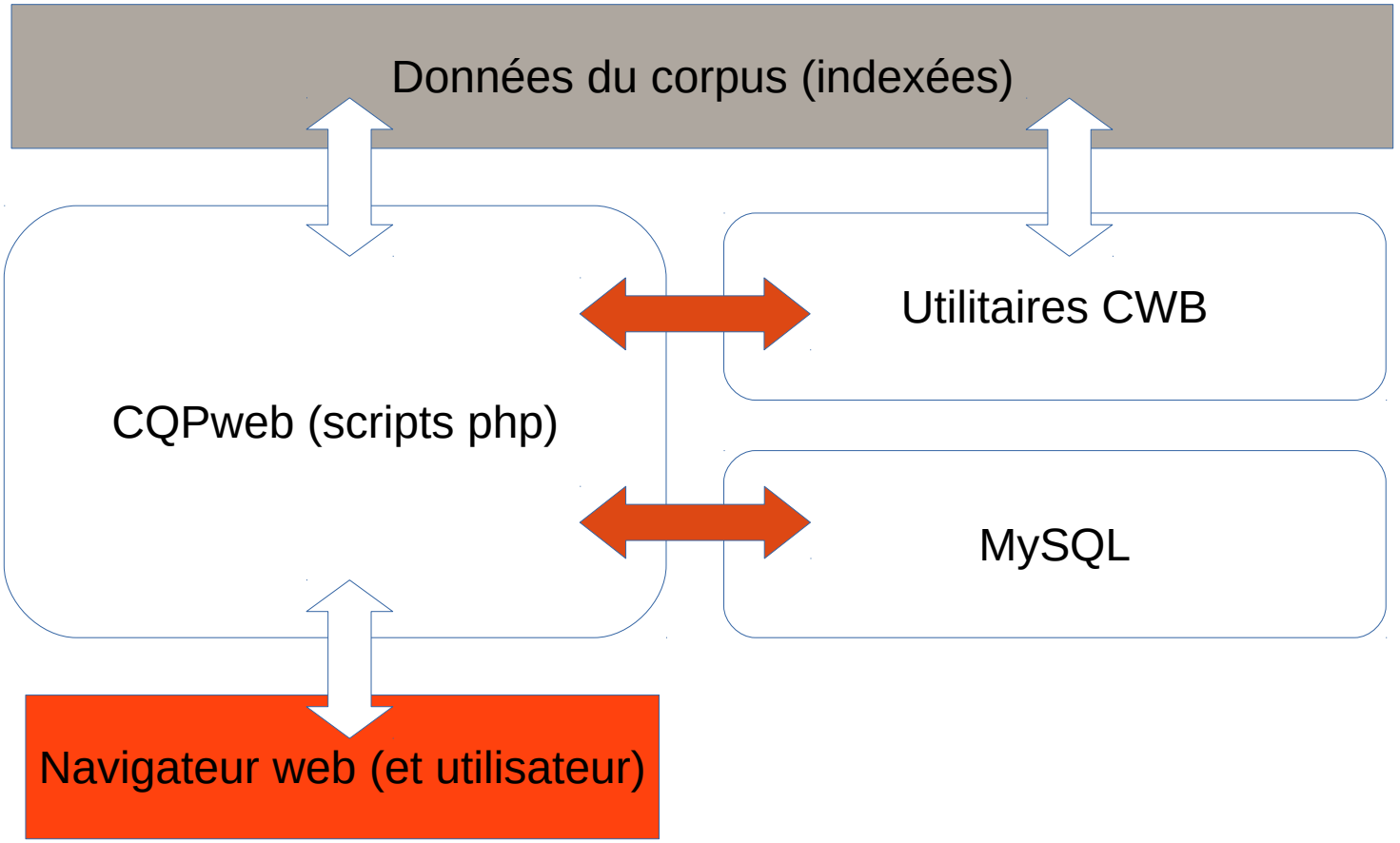
Inconvénients : lié par son architecture à un seul corpus.

cqpweb.univ-avignon.fr

cqpweb : l’architecture

Réécriture de BNCweb, en php, qui pourra servir d'interface pour n'importe quel corpus, à condition qu’il soit convenablement préparé (indexé).

cqpweb.univ-avignon.fr



cqpweb.univ-avignon.fr

cqpweb : utilisation … et utilité

- Requêtes simples : mots et suites de mots ;
- Requêtes par POS ou par lemme, utilisation de caractères génériques ;
- Combinaisons de ces différents types de requêtes

cqpweb.univ-avignon.fr

cqpweb: utilisation … et utilité

- Requêtes simples : mots et suites de mots :

« saisir », « fût -il », « on aurait dit »

cqpweb.univ-avignon.fr

cqpweb : utilisation … et utilité

- Requêtes par POS ou par lemme

_ADJ : tout adjectif
{saisir} : toutes les formes du verbe « saisir »

cqpweb.univ-avignon.fr

cqpweb : utilisation … et utilité

- Combinaisons de ces différents types de requêtes, caractères génériques, alternatives…

{saisir} * _NOM : toute forme de « saisir » suivie facultativement d’une suite de caractères suivie d’un nom …

cqpweb.univ-avignon.fr

cqpweb : utilisation … et utilité

- Option de syntaxe cqp qui permet des requêtes plus complexes

n1:[pos = "NOM"] [pos = "PRP"] n2:[pos = "NOM"]:: n1.lemma = n2.lemma

→ *tête à tête*, *service pour service*, *jour en jour*, etc.

cqpweb.univ-avignon.fr

cqpweb : utilisation … et utilité

Le corpus est composé d’un nombre de textes, qui peuvent être étiquetés pour POS et lemmes, par exemple, et auxquels peuvent être associées des catégories TEI ou *ad hoc* (auteur, date, genre, etc.)

cqpweb.univ-avignon.fr

cqpweb : utilisation … et utilité

Les résultats des requêtes peuvent être triés par fréquence, par distribution (dans quels textes ? quelles catégories ? etc.), ou bien interrogés pour les collocations avec d’autres items (ex. {prendre} + NOM).

cqpweb.univ-avignon.fr

cqpweb : utilisation … et utilité

Les résultats peuvent également être annotés par l'utilisateur, selon ses propres critères (ex. « saisir » avec compléments concrets ou abstraits) ; ses résultats sont sauvegardés dans sa session, ou peuvent être téléchargés pour une utilisation hors ligne.

cqpweb.univ-avignon.fr

… et l’interdisciplinaire

cqpweb se prête à une utilisation interdisciplinaire dès lors qu’on mène une recherche susceptible de s’intéresser au repérage de régularités textuelles (potentiellement multivariées).

L’interface est rustique, mais néanmoins claire, permettant aux non-informaticiens ou non-statisticiens de formuler des requêtes complexes.

cqpweb.univ-avignon.fr

… et l’interdisciplinaire

Ces régularités linguistiques peuvent renseigner sur la langue elle-même (constructions, lexèmes, collocations, par période, par auteur, par genre, etc.)

Elles peuvent également renseigner sur le type de texte en jeu : style, lexique, thématiques, constructions, etc.

cqpweb.univ-avignon.fr

… et l’interdisciplinaire

Exploitations possibles donc en littérature, en SHS ou en droit, mais également des les sciences dures, dès lors qu’il s’agit de travailler sur des données textuelles dont le volume se prête à ce type d’analyse par algorithmes.

cqpweb.univ-avignon.fr

… et l’interdisciplinaire

L’interface à l’UAPV ne comprend, pour l’instant, que deux corpus : un corpus de démonstration et une version bêta du corpus des étudiants de Master

(Pour essayer : login astudent, mdp astudent.)

L’interface ici : https://cqpweb.lancs.ac.uk/ montre cependant l’étendue des possibilités que présente cette architecture.

cqpweb.univ-avignon.fr

Quelques références :

Evert, Stefan & Andrew Hardie. 2011. « Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. » Presentation at Corpus Linguistics 2011, University of Birmingham, UK.
Hardie, Andrew. 2012. « CQPweb — combining power, flexibility and usability in a corpus analysis tool. » International Journal of Corpus Linguistics 17(3). 380–409. doi:10.1075/ijcl.17.3.04har.
Hardie, Andrew. 2014. « Modest XML for Corpora: Not a standard, but a suggestion. » ICAME Journal 38(1). doi:10.2478/icame-2014-0004. http://www.degruyter.com/view/j/icame.2014.38.issue-1/icame-2014-0004/icame-2014-0004.xml (19 December, 2015).
Hoffmann, Sebastian. 2008. Corpus linguistics with BNCweb: a practical guide. (English Corpus Linguistics v. 6). Frankfurt am Main: Peter Lang.