Un corpus réservoir d'exemples lexicographiques

Jean-Luc Benoit Veronika Lux-Pogodalla

Mai 2012

Spécification initiale

- un réservoir d'exemples lexicographiques pour RLF
- composé de textes d'une certaine variété
- de taille importante : au moins un milliard de mots
- ~~un corpus représentatif de la langue française~~

Chantiers ouverts

- recensement pragmatique des ressources disponibles
- identification de "manques" par rapport à un "idéal"
- recherche d'un outil de consultation de ce corpus

Ressources disponibles à l'ATILF

- Existant :
  - Frantext, L'Est Républicain
  - Quelques centaines de milliers de mots
- Manquent :
  - Oral, PQN, écrits scientifiques et techniques, ouvrages pratiques, modes d'emploi, blogs, presse enfantine, publicité, presse d'entreprise et syndicale...
  - Encore quelques mots pour arriver au milliard !

Ressources disponibles hors ATILF : FrWac

Caractéristiques :

- tranche de Web (1,6 milliards de mots) constituée de pages sélectionnées et étiquetées (TreeTagger)
- 16 fichiers XML préparés pour indexation/interrogation avec CorpusWorkBench ou CQPWeb
- constitué dans le cadre de WaCky (corpus équivalents pour anglais, allemand, italien, français)

Avantages : gros corpus pérenne illustrant la variété du Web
Inconvénients : composition exacte inconnue et nombreuses pages inadéquates pour chercher des exemples lexicographiques
Amélioration en cours : ajout à chaque page de FrWac d'un indicateur de qualité basé sur les analyses linguistiques fournies par Antidote-Druide

Fonctionnalités attendues

- De base : concordancier
- Avancées (reposant sur un enrichissement des données) :
  - recherche de cooccurrences
  - recherche sur un vocable
  - recherche sur une catégorie grammaticale
  - etc.

État de l'art

- Nombreux outils :
  - gratuits ou pas
  - fonctionnant avec un corpus unique ou pas
  - maintenus ou pas
  - etc.
- Exploration approfondie de :
  - Frantext/Stella : interface d'interrogation de Frantext
  - CQPWeb : interface d'interrogation de l'Est Républicain et de FrWac

Bilan

- Frantext/Stella et CQPWeb sont des concordanciers avec quelques fonctionnalités avancées (recherche cooccurrences, recherche sur les lemmes, recherche sur catégorie grammaticale).
- Frantext/Stella :
  - + : interface assez riche et expertise interne sur le langage de requête
  - - : difficulté à monter de nouveaux textes, ergonomie désuète pour certains, petite sous-partie de Frantext étiquetée en parties du discours
- CQPWeb
  - + : montage de nouveaux textes assez facile (XML) mais richesse de l'interrogation dépendante des méta-données disponibles
  - - : langage de requête relativement limité

Conclusion

- Corpus en chantier
- Difficultés informatiques inédites liées au volume de FrWac
- ... et pourtant, on en veut encore plus (ex. PQN) !