Un nouveau « GlossaNet » : recherche d'exemples linguistiques dans des corpus issus du Web

C. Fairon, H. Naets, K. Macé

GlossaNet est un service en ligne gratuit qui permet de faire des recherches d'occurrences linguistiques dans des corpus « dynamiques » issus du Web.

Afin d'utiliser cet environnement, les utilisateurs :

- créent un compte personnel sur le site Web
- choisissent une série de sources Internet pour constituer un « corpus dynamique » ;
- et enregistrent une ou plusieurs « requêtes linguistiques » sur ce corpus.

Le système GlossaNet prend en charge le téléchargement des corpus et vérifie régulièrement si de nouveaux textes sont disponibles dans les sources référencées. À chaque fois que le corpus est mis à jour, les requêtes des utilisateurs sont réappliquées et les nouveaux résultats sont transmis sous forme de concordance à l'utilisateur.

Ancienne version

Né en 1998 à l'Université de Paris 7, GlossaNet permettait de faire des recherches dans une centaine de journaux représentant une dizaine de langues. Le succès rencontré par l'application a été assez important puisque, dans les derniers mois, quelque 2300 utilisateurs étaient enregistrés, tandis que le système traitait de l'ordre de 30.000 requêtes par jour.

À l'exception d'outils tels que WebCorp[1] ou WebConcordancer[2] qui traitent le Web comme un corpus, les nombreux systèmes de concordances disponibles en ligne donnent en général accès à des corpus « figés ». GlossaNet, au contraire, offre au linguiste la possibilité d'obtenir continuellement de nouvelles concordances dans la mesure où le corpus change au cours du temps et où les requêtes sont automatiquement réappliquées. En outre, les textes traités par GlossaNet font l'objet d'une analyse linguistique, ce qui permet aux utilisateurs de réaliser des requêtes complexes intégrant des formes fléchies, des lemmes, des parties du discours, ainsi que des informations morphologiques ou même sémantiques.

Nouvelle version

Nous présenterons dans cette démonstration la **nouvelle version de GlossaNet** lancée en mai 2008. Cette nouvelle version permet aux utilisateurs de composer leurs corpus en spécifiant une série de sources ATOM/RSS. Le champ d'action de GlossaNet n'est donc plus limité à un ensemble pré-sélectionné de sources, mais est ouvert à toute source publiée sous forme de RSS (ce qui est très fréquemment le cas des journaux, blogs, forums, etc.). Un autre apport de cette approche est que les Flux RSS fournissent en général des articles catégorisés : il est donc facile de constituer des corpus spécialisés par thème (sport, santé, médecine, politique, culture, etc.) ou par genre de textes (argumentatif, informatif, discussions informelles, etc).

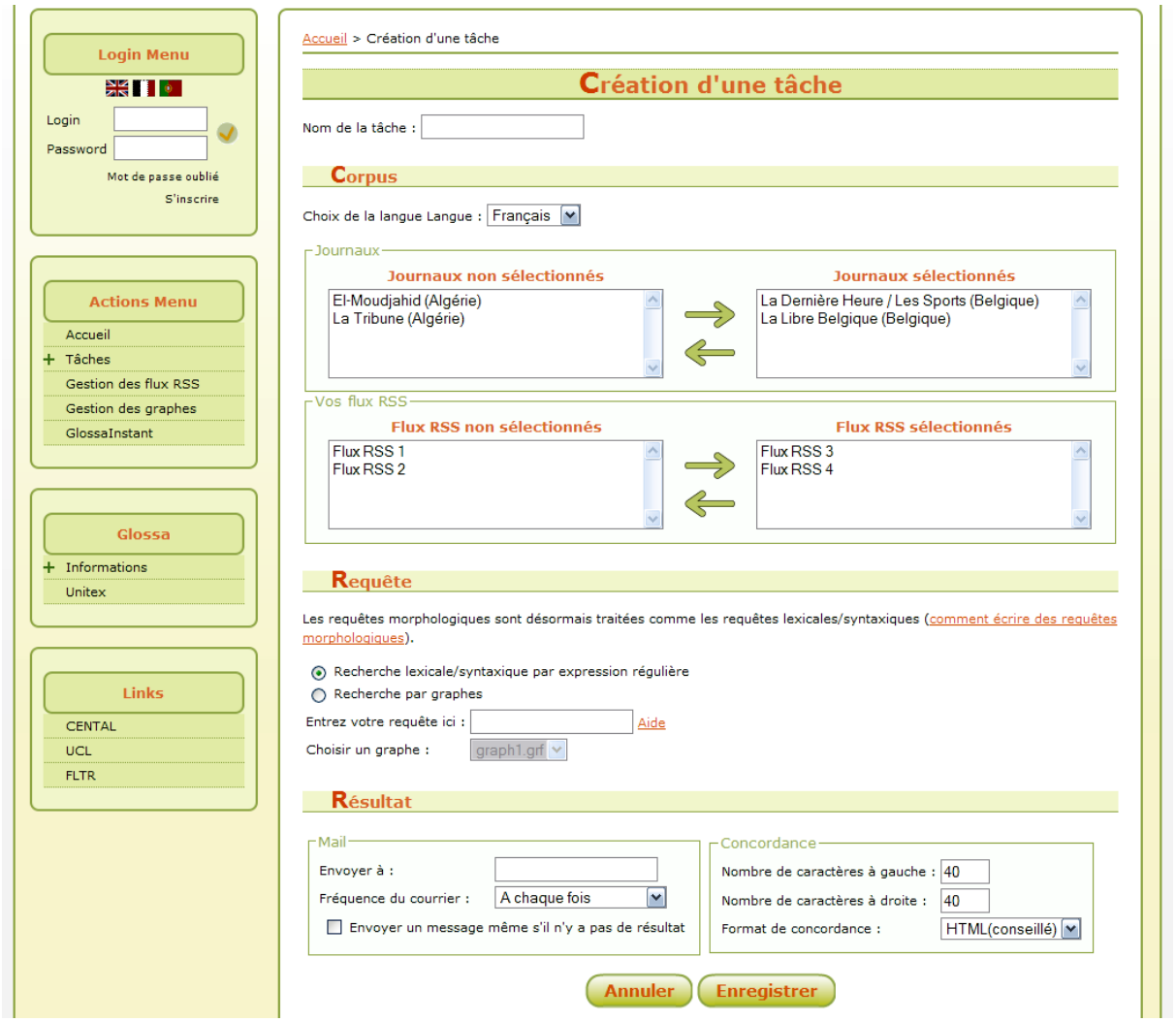
---

[1] http://www.webcorp.org.uk/

[2] http://webascorpus.org/searchwac.html

GlossaNet intègre deux logiciels existant par ailleurs : Unitex (Paumier 2003), un logiciel d'analyse de corpus disposant de dictionnaires électroniques à large couverture, et Corporator (Fairon 2006), un logiciel qui permet de constituer des corpus par téléchargement de flux ATOM/RSS.

GlossaNet a été conçu pour les linguistes qui ont souvent besoin de larges corpus pour trouver des occurrences pertinentes pour leurs recherches. Ici, les corpus changent avec le temps… Plus on attend, plus on reçoit d'exemples. L'expérience a par ailleurs montré que certains utilisateurs ont recours au système GlossaNet pour chercher des mots clés qui traduisent plus des préoccupations en matière de veille d'information (type revue de presse automatique) qu'en matière linguistique.



Demande technique :

Connexion internet et table + espace pour poster.