Scientext

Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques

Achille Falaise (LIG-GETALP), Agnès Tutin (LIDILEM), Olivier Kraif (LIDILEM)

Situation

- Des corpus
  - Annotés : lemme, partie du discours, flexion
  - Arborés (*treebanks*) : relations syntaxiques
  - Structurés : type de texte, partie textuelle
- Des outils
  - Peu nombreux (pour les corpus arborés)
  - Basés sur des expressions régulières, complexes pour des non-informaticiens (linguistes, didacticiens, etc.)

```
<cat=V,#1> && <lemma=hypothèse,#2> :: (OBJ,#2,#1)
```

Exemple : langage de requête de *ConcQuest*

Scientext

- Un environnement en ligne pour non-informaticiens
- Développé au-dessus du moteur de recherche ConcQuest

Corpus librement consultables (projet Scientext → Scientext 1.3)

Type	Langue	Analyseur	Nb mots
Publications scientifiques	français	Syntex	5M
Publications scientifiques	anglais	Syntex	14M
Mémoires d'apprenants	anglais	Syntex	1M

Corpus à accès restreint (projet Emolex → Scientext 1.4)

Type	Langue	Analyseur	Nb mots
Textes littéraires	français	Connexor	66M
Textes journalistiques	français	Connexor	226M
Textes journalistiques	russe	TreeTagger + DeSR / SynTagRus	0,5M

Extension en cours pour des corpus allemands et espagnols (Connexor), et anglais (XIP)

Recherche dans un corpus en 5 étapes

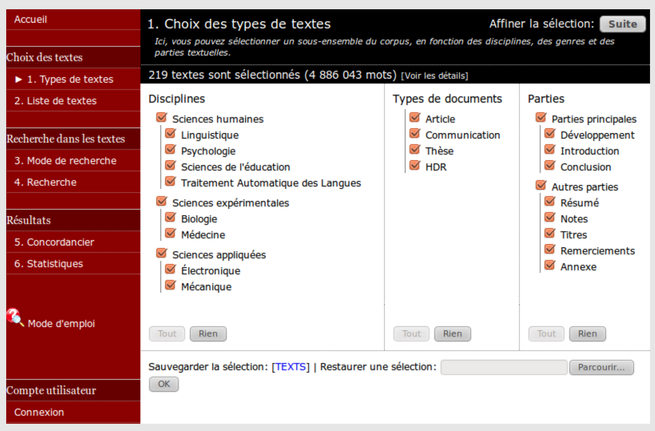
1 Choix d'un corpus

2 Sélection d'un sous-corpus

On commence par sélectionner les parties du corpus que l'on souhaite étudier, en fonction de la structure du corpus.

Une sélection fine (texte par texte) est possible dans un autre écran.

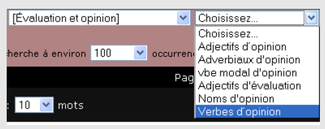
La liste des parties sélectionnées peut être sauvegardée.



3 Création d'une requête

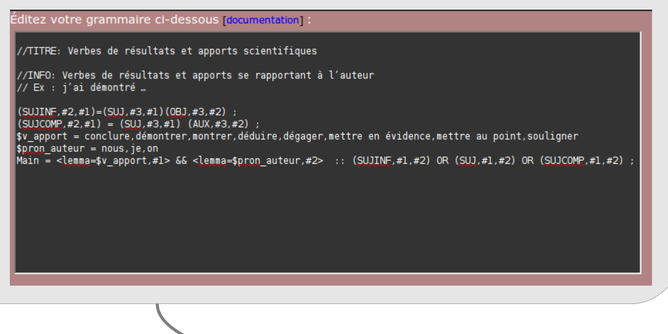
3 modes sont disponibles, suivant les besoins et les capacités de l'utilisateur.

Recherche sémantique :
on sélectionne une requête prédéfinie, créée au préalable par l'équipe.



Recherche libre :
on compose une requête à l'aide d'un assistant, dévoilant progressivement ses fonctionnalités.

Recherche avancée :
on compose une requête à l'aide d'un langage de grammaires spécialisé, permettant l'utilisation de listes, la définition de relations syntaxiques profondes, etc.



4 Visualisation et contrôle des résultats

On peut ensuite :

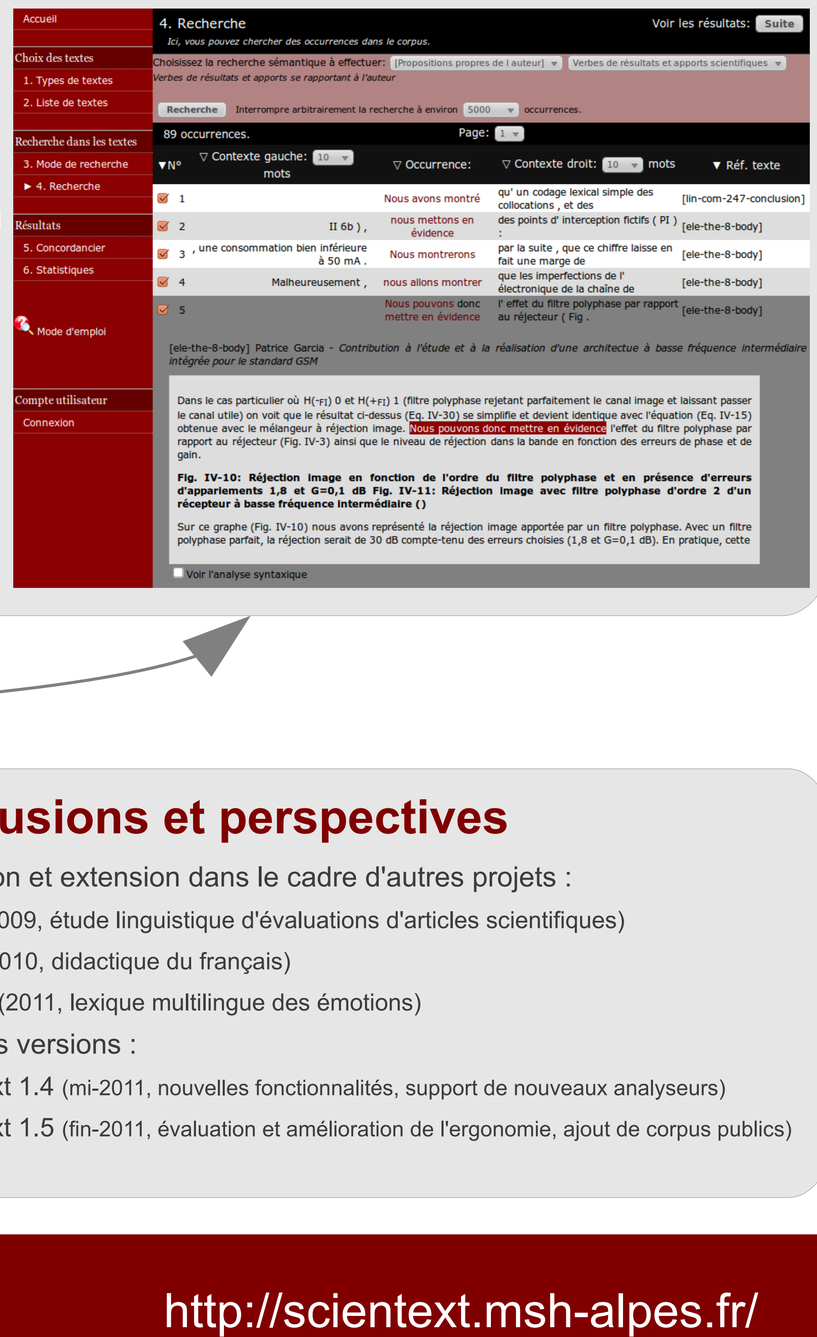
Consulter les résultats dans un affichage *KWIC* (KeyWord In Context).

Élargir le contexte de chaque résultat (la mise en forme du texte original est préservée dans une certaine mesure).

Visualiser l'analyse de la phrase.

Désactiver les résultats incorrects.

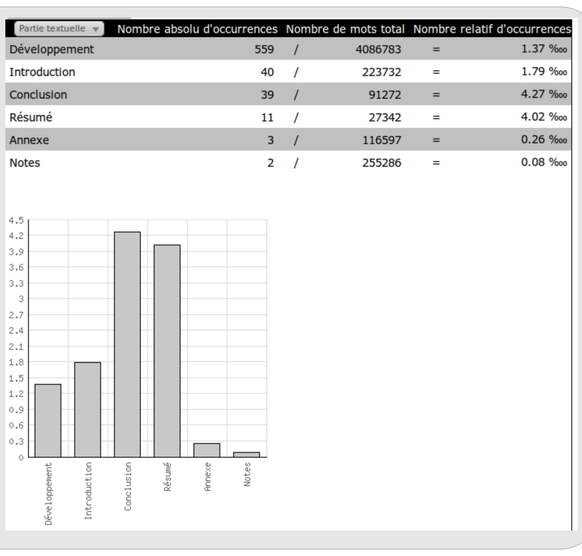
Exporter les résultats au format CSV ou XLS.



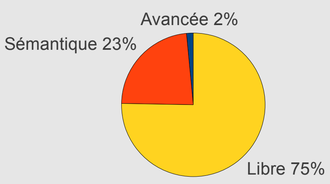
5 Statistiques

Enfin, on peut consulter des statistiques correspondant aux résultats : fréquence des lemmes, répartition par type de texte, etc., en fonction de la structure du corpus.

Ces statistiques peuvent être exportées au format CSV ou XLS.

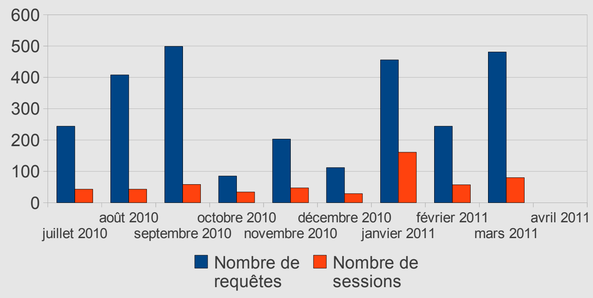


Statistiques d'utilisation



Répartition des requêtes par mode

Entre juillet 2010 et mars 2011, 4432 requêtes ont été effectuées (en 662 sessions). La recherche libre est utilisée pour 75% des requêtes, la recherche sémantique pour 23% et la recherche avancée pour 2%.



Fréquentation mensuelle de l'environnement Scientext, pour les corpus publics, de juillet 2010 à mars 2011

Conclusions et perspectives

Réutilisation et extension dans le cadre d'autres projets :

- Cedill (2009, étude linguistique d'évaluations d'articles scientifiques)
- FULS (2010, didactique du français)
- Emolex (2011, lexique multilingue des émotions)

Prochaines versions :

- Scientext 1.4 (mi-2011, nouvelles fonctionnalités, support de nouveaux analyseurs)
- Scientext 1.5 (fin-2011, évaluation et amélioration de l'ergonomie, ajout de corpus publics)

Agence Nationale de la Recherche ANR
LIDILEM Université Stendhal Grenoble
Laboratoire Langage Littérature Société
UBS Université de Bretagne-Sud

http://scientext.msh-alpes.fr/