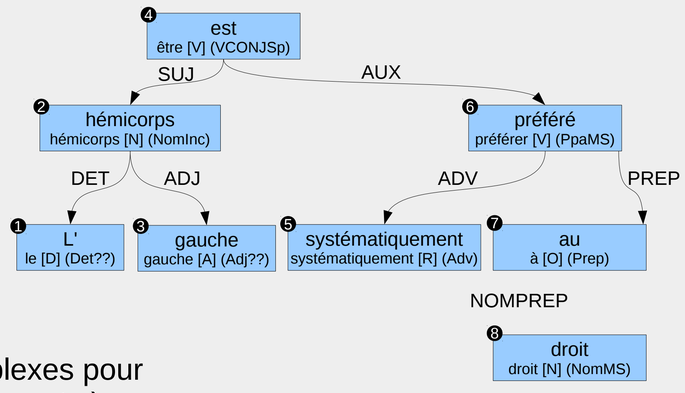
La plateforme ScienQuest

Achille Falaise

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*

ScienQuest

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

Corpus Scientext (Agnès Tutin, Francis Grossmann)

Type	Langue	Analyseur	Nb mots
Publications scientifiques	français	Syntex	5M
Évaluations de communications	français	Syntex	34k
Publications scientifiques	anglais	Syntex	14M
Mémoires d'apprenants	anglais	Syntex	1M

Corpus TALN (Florian Boudin, Ludovic Tanguy)

Type	Langue	Analyseur	Nb mots
Textes scientifiques	français	Connexor	2M

Corpus Est Républicain (Bertrand Gaiffe, Kamel Nehbi)

Type	Langue	Analyseur	Nb mots
Articles de presse	français	MElt + Malt	110M

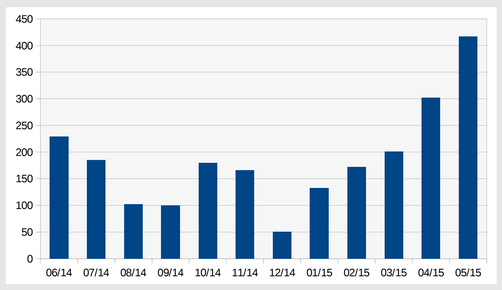
Corpus Scientext – Métadonnées :
(publications scientifiques en français)
- Discipline (linguistique, biologie, etc.)
- Genre (article, thèse, etc.)
- Partie (développement, introduction, etc.)

Corpus TALN – Métadonnées :
- Conférence (RÉCITAL, TALN)
- Année (2007-2013)
- Type (long, court, etc.)

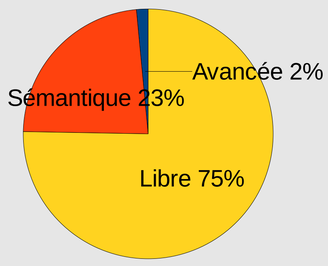
Corpus Est Républicain :
- Année (1999, 2002, 2003)
- Partie (accroche, article)

Usage

Actuellement, ScienQuest est principalement utilisé par des linguistes, pour de la linguistique de corpus.



Nombre de sessions par mois (juin 2014 – mai 2015)



Répartition des requêtes par mode (2010-2011)

Perspectives : corpus pour non spécialistes

Aide à la rédaction en français :

- TUTIN A., FALAISE A. (2013). Multiword expressions in scientific discourse: a corpusdriven database. In *Proceedings of the 3rd biennial conference on electronic lexicography 2013*, Tallinn, Estonie.

Aide à la rédaction en anglais :

- JACQUES M.P., HARTWELL L., FALAISE A. (2013). TAL et linguistique de corpus pour aider la rédaction scientifique en anglais. In *Actes de Traitement Automatique du Langage Naturel 2013*, Les Sables d'Olonne, France.

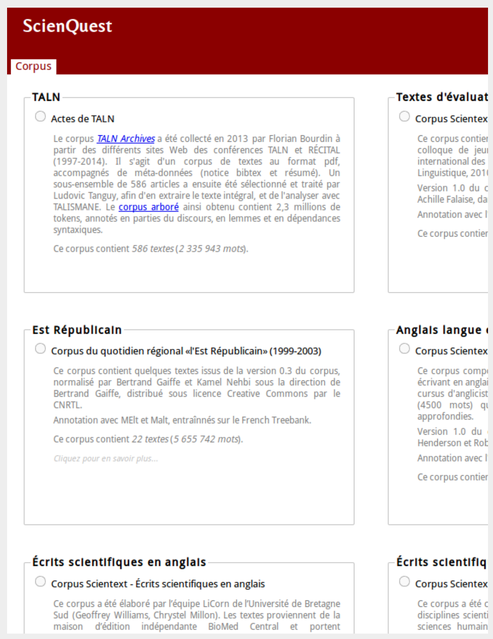
Didactique des langues (projet Dicorpus) :

- ROSSI C., FRÉROT C., FALAISE A. (2014). Integrating controlled corpus data in the classroom: a case-study of English NPs for French students in specialised translation. In *Proceedings of the 6th International Conference on Corpus Linguistics*, Las Palmas de Gran Canaria, Espagne.

Recherche dans un corpus en 5 étapes

1 Choix d'un corpus

On commence par sélectionner un corpus parmi les corpus arborés disponibles.

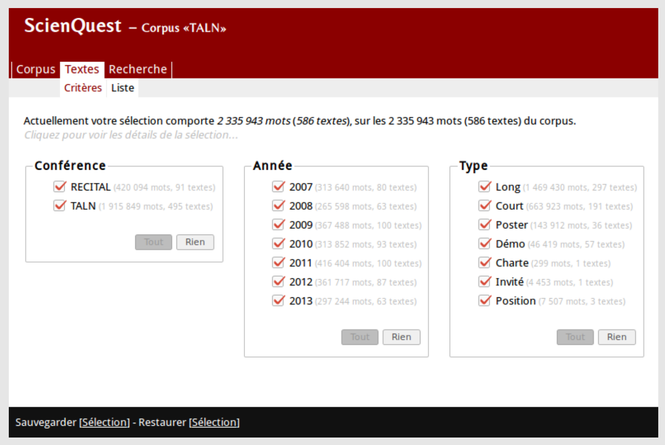


2 Sélection d'un sous-corpus

Puis on sélectionne les parties du corpus que l'on souhaite étudier, en fonction des métadonnées disponibles dans le corpus.

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

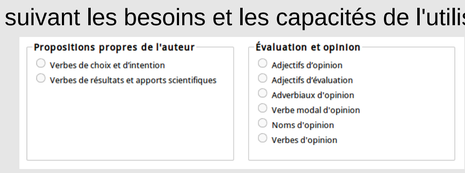
La liste des parties sélectionnées peut être sauvegardée.



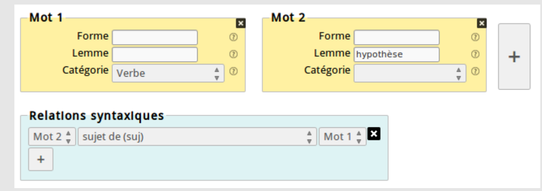
3 Création d'une requête

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

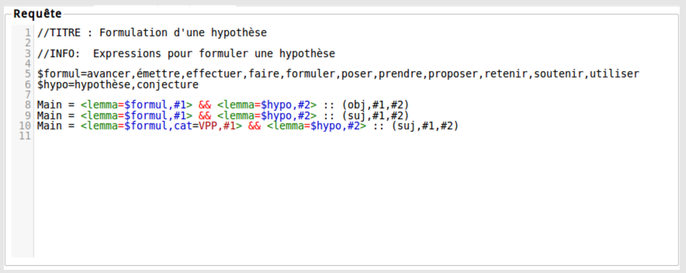
Recherche sémantique :
on sélectionne une requête prédéfinie, créée au préalable par l'équipe.



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 syntaxique de la phrase.

Désactiver les résultats incorrects.

Exporter les résultats au format CSV, XLS ou SQR2.

	#	Contexte gauche	Occurrences	Contexte droit	Texte
☑	1	site web . Nous verrons dans la discussion que cette	hypothèse se vérifie	lors_de l' étude des performances sur les différents corpus .	#9 - RECITAL - Poster - Article
☑	2	de m et n (2) . Même_si ces	hypothèses paraissent	légitimes , les systèmes d' alignement se basant dessus rencontrent	#11 - RECITAL - Poster - Article
☑	3	le résultat final obtenu . Or des traitements , des	hypothèses sont instaurés	à différentes étapes de la résolution des questions . Nous	#12 - RECITAL - Poster - Article
☑	4	Prize in 1989 ? serait Nobel Peace Prize car l'	hypothèse est faite	que la réponse correcte devrait se trouver la proximité de	#12 - RECITAL - Poster - Article
☑	5	en_fonction_des thèmes qui sont abordés dans leurs parties . Notre	hypothèse est	que les ruptures thématiques fournies par cet algorithme peuvent devenir	#13 - RECITAL - Poster - Article
☑	6	textes fondée sur la notion de changement thématique . L'	hypothèse de Hearst (in Hernandez , 2004 : 191 , note	n° 86) est que " un ensemble d' items	#13 - RECITAL - Poster - Article
☑	7	parties titrées courtes (de niveau 3) . L'	hypothèse selon laquelle les frontières de segments thématiques peuvent servir d' indices contribuant au repérage des segments d' information évolutive semble	se confirmer même_s' il s' avère nécessaire d' approfondir les	#13 - RECITAL - Poster - Article
☑	8	de termes ambigüs apparaissant dans un nouveau texte . Cette	hypothèse sera expérimentée	très prochainement dans la suite de ces travaux . Par_ailleurs	#33 - RECITAL - Long - Article

Afficher : ○ Contexte ○ Arbre syntaxique

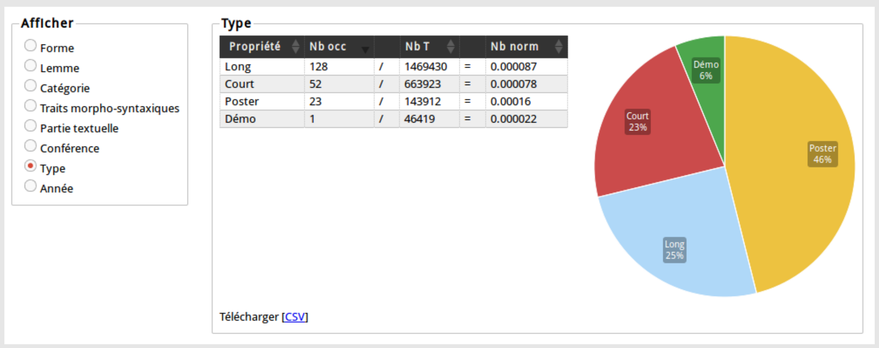
pas de méthode pour apprendre les paramètres optimaux et nous restons ainsi contraints de les choisir par l' expérimentation . Un avantage des méthodes présentées peut être mis en_avant : nous supposons que le fait d' utiliser les plus proches voisins comme ensemble d' éléments à regrouper (et non les cooccurrents) , permettra d' utiliser ces voisins comme données d' apprentissage pour un classifieur destiné à la désambiguïsation de termes ambigüs apparaissant dans un nouveau texte . Cette hypothèse sera expérimentée très prochainement dans la suite de ces travaux . Par_ailleurs , la finalisation de ces travaux nécessite diverses études à mener rapidement . Afin_de mettre en valeur les apports de la méthode présentée ici , nous souhaitons former des clusters à_partir_des mêmes ensembles de plus proches voisins que dans cette étude mais en utilisant d'une_part l' espace de cooccurrences de fenêtres seul ainsi_que chacun des espaces syntaxiques seul , et en utilisant d'autre_part les

2009-long
recital-2009-long-005
2009

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.



ENS DE LYON
icar UMR 5191 - CNRS
LIDILEM Université Stendhal Grenoble 3

http://corpora.aiakide.net