La plateforme ScienQuest Achille Falaise

Corpus Scientext – Métadonnées :

• Discipline (linguistique, biologie, etc.)

• Partie (développement, introduction, etc.)

• Type (long, court, etc.)

Corpus Est Républicain :

• Année (1999, 2002, 2003)

• Partie (accroche, article)

(publications scientifiques en français)

• Genre (article, thèse, etc.)

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*



Recherche dans un corpus en 5 étapes

Choix	ScienQuest Corpus	
d'un cornus		Textes d'évaluat
On commence par sélectionner un corpus parmi les corpus arborés disponibles.	Le corpus <u>TALN Archives</u> a été collecté en 2013 par Florian Bourdin à partir des différents sites Web des conférences TALN et RÉCITAL (1997-2014). Il s'agit d'un corpus de textes au format pdf, accompagnés de méta-données (notice bibtex et résumé). Un sous-ensemble de 586 articles a ensuite été sélectionné et traité par Ludovic Tanguy, afin d'en extraire le texte intégral, et de l'analyser avec TALISMANE. Le <u>corpus arboré</u> ainsi obtenu contient 2,3 millions de tokens, annotés en parties du discours, en lemmes et en dépendances syntaxiques. Ce corpus contient <i>586 textes (2 335 943 mots</i>).	Ce corpus outlier colloque de jeu international des Linguistique, 201 Version 1.0 du o Achille Falaise, da Annotation avec l Ce corpus contier
	Est Républicain	Anglais langue
	Corpus du quotidien régional «l'Est Républicain» (1999-2003)	Corpus Scientex
	Ce corpus contient quelques textes issus de la version 0.3 du corpus, normalisé par Bertrand Gaiffe et Kamel Nehbi sous la direction de Bertrand Gaiffe, distribué sous licence Creative Commons par le CNRTL. Annotation avec MElt et Malt, entraînnés sur le French Treebank.	Ce corpus comp écrivant en angla cursus d'anglicis (4500 mots) q approfondies.
	Ce corpus contient 22 textes (5 655 742 mots).	Version 1.0 d Henderson et R

–Écrits scientifiques en anglais Écrits scientifiq Corpus Scientext - Écrits scientifiques en anglais

Ce corpus a été élaboré par l'équipe LiCorn de l'Université de Bretagne Sud (Geoffrey Williams, Chrystel Millon). Les textes proviennent de la maison d'édition indépendante BioMed Central et portent

Corpus Scientex Ce corpus a été c disciplines scienti sciences humain

Annotation avec Ce corpus contie

ScienQuest

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

Corpus Scientext (Agnès Tutin, Francis Grossmann)

Туре	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)

IPUS IALIN (Fiorian Boud	Corpus TALN – Métadonnées :			
Туре	Langue	Analyseur	Nb mots	Conférence (RÉCITAL, TALN)
Textes scientifiques	français	Connexor	2M	 Année (2007-2013)

Corpus Est Républicain (Bertrand Gaiffe, Kamel Nehbi)

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

2 Sélection d'un sous-corpus

Puis on sélectionne les ScienQuest - Corpus «TALN» parties du corpus que l'on souhaite étudier, Corpus Textes Recherche Critères Liste en fonction des métadonnées Actuellement votre sélection comporte 2 335 943 mots (586 textes), sur les 2 335 943 mots (586 textes) du corpus. disponibles dans le Conférence Année corpus. 2007 (313 640 mots, 80 textes) 🗹 Long (1 469 430 mots, 297 texte RECITAL (4) **TALN** (1 915 849 mots, 495 textes) 2008 (265 598 mots, 63 textes) Court (663 923 mots, 191 texter Une sélection fine 2009 (367 488 mots, 100 textes) Poster (143 912 mots, 36 textes) (texte par texte) est Tout Rien 2010 (313 852 mots, 93 textes) Démo (46 419 mots, 57 textes Charte (299 mots, 1 textes) possible dans un autre 2011 (416 404 mots, 100 textes 2012 (361 717 mots, 87 textes) Invité (4 453 mots, 1 textes) écran. 2013 (297 244 mots, 63 textes Position (7 507 mots, 3 textes) La liste des parties Tout Rien Tout Rien sélectionnées peut être sauvegardée. vegarder [<u>Sélection</u>] - Restaurer [<u>Sélection</u>]

Création d'une requête

Usage

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





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 :

5

corpus.

• 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 casestudy 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.

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

Propositions propres de l'auteur

Verbes de résultats et apports scientifiques

erbes de choix et d'intention

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

on compose une requête à

Recherche libre :

progressivement ses

fonctionnalités.

Mot 1		E	ĸ	Mot 2			×	
Forme		0		Forme			0	
Lemme		0		Lemme	hypothèse		0	+
Catégorie	Verbe	▲ ▼		Catégorie		≜ ⊽	0	

Évaluation et opini Adjectifs d'opinion

Adjectifs d'évaluation

Adverbiaux d'opinio

Verbe modal d'opinio

Noms d'opinion Verbes d'opinion

l'aide d'un assistant, dévoilant Relations syntaxiques Mot 2 🛔 sujet de (suj) 🛔 Mot 1 🛔 👗 +

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.

Requête //TITRE : Formulation d'une hypothèse //INF0: Expressions pour formuler une hypothèse \$formul=avancer, émettre, effectuer, faire, formuler, poser, prendre, proposer, retenir, soutenir, utiliser \$hypo=hypothèse,conjecture

Main = <lemma=\$formul,#1> && <lemma=\$hypo,#2> :: (obj,#1,#2)
Main = <lemma=\$formul,#1> && <lemma=\$hypo,#2> :: (suj,#1,#2)
Main = <lemma=\$formul,cat=VPP,#1> && <lemma=\$hypo,#2> :: (suj,#1,#2)

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 gau	iche	Occurrences	Contexte droit	Texte
✓	1	site web . Nous verron	is dans la discussion que cette	hypothèse se vérifie	lors_de l' étude des performances sur les différents corpus .	#9 - RECITAL Poster - Artic
✓	2	de m et	n (2) . Même_si ces	hypothèses paraissent	légitimes , les systèmes d' alignement se basant dessus rencontrent	#11 - RECITA Poster - Artic
~	3	le résultat final obtenu .	, Or des traitements des	hypothèses sont instaurés	à différentes étapes de la résolution des questions . Nous	#12 - RECITA Poster - Arti
✓	4	Prize in 1989 ? serait No	bel Peace Prize car l'	hypothèse est faite	que la réponse correcte devrait se trouver la proximité de	#12 - RECITA Poster - Arti
~	5	en_fonction_des thèm dans	es qui sont abordés leurs parties . Notre	hypothèse est	que les ruptures thématiques fournies par cet algorithme peuvent devenir	#13 - RECITA Poster - Arti
✓	6	textes fondée sur la no	tion de changement thématique . L'	<mark>hypothèse</mark> de Hearst (in Hernandez , 2004 : 191 , <mark>note</mark>	n° 86) est que " un ensemble d' items	#13 - RECITA Poster - Arti
✓	7	parties titrées court	tes (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 - RECIT/ Poster - Arti
✓	8	de termes ambigüs aj n	pparaissant dans un ouveau texte . Cette	hypothèse sera expérimentée	très prochainement dans la suite de ces travaux . Par_ailleurs	#33 - RECIT/ Long - Articl
- A1	f iche Conte Arbre	exte e syntaxique	pas de méthode p expérimentation . utiliser les plus pro utiliser ces voisins ambigüs apparaise suite de ces travau Afin_de mettre en mêmes ensembles	our apprendre les paramètres optimaux et nous restons ainsi contraints de les ch Un avantage des méthodes présentées peut être mis en_avant : nous supposons oches voisins comme ensemble d' éléments à regrouper (et non les cooccurrents) comme données d' apprentissage pour un classifieur destiné à la désambiguïsatio sant dans un nouveau texte . Cette hypothèse sera expérimentée très prochaine ux . Par_ailleurs , la finalisation de ces travaux nécessite diverses études à mener ra valeur les apports de la méthode présentée ici , nous souhaitons former des cluste s de plus proches voisins que dans cette étude mais en utilisant d'une_part l' espace	oisir par l' que le fait d'), permettra d' 2009-long <i>recital-2009-long-005</i> 2009 on de termes ement dans la pidement . ers à_partir_des ce de	

http://corpora.aiakide.net



