Approche onomasiologique de la phraséologie transdisciplinaire des écrits scientifiques : la recherche sémantique dans les textes dans le cadre du projet Scientext

Falaise Achille[+], Tutin Agnès[*]
+Equipe GETALP, LIG, Grenoble :
achille.falaise@imag.fr
Laboratoire LIDILEM, Université Grenoble 3-Stendhal : *agnes.tutin@u-grenoble3.fr

Résumé : L'accès à la phraséologie, en particulier pour les applications en langue étrangère et seconde, se fait rarement à partir de corpus. Dans le cadre du projet ANR Scientext, nous avons élaboré un mode d'accès à la phraséologie transdisciplinaire des écrits scientifiques par un mode onomasiologique. Des requêtes prédéfinies portant sur la question linguistique du positionnement et du raisonnement ont été élaborées à partir de schémas syntaxiques et sémantiques, par exemple sur l'expression de l'opinion ou de l'évaluation. Ces grammaires sont ensuite appliquées à un large corpus d'écrits scientifiques balisé au plan structurel et au plan syntaxique (analyse de dépendance). L'utilisateur peut ainsi extraire, selon ses besoins, une phraséologie adaptée à une requête sémantique.

Point matériel : Les auteurs apporteront leurs ordinateurs portables. Une plaquette du projet pourra être distribuée.

1. Introduction

Dans cette démonstration, nous souhaitons présenter les modes d'accès aux informations lexicales et phraséologiques élaborés dans le cadre du projet ANR Scientext 2007-2010 "Corpus et outils de la recherche en sciences humaines et sociales" que nous pilotons au LIDILEM (U-Grenoble 3).

Dans ce cadre, un site web (*www.u-grenoble3.fr/lidilem/scientext*), qui permet un accès aux écrits scientifiques, a été élaboré. Nous souhaitons mettre l'accent dans cette démonstration sur l'accès sémantique à la phraséologie des écrits scientifiques qui constitue une originalité de notre projet (Tutin, à paraître).

2. Trois modes d'accès aux textes

2.1 Le corpus

Dans le cadre de ce projet, un large corpus d'écrits scientifiques, a été constitué[1]. A ce jour, il contient 4,3 millions de mots dans des disciplines variées (linguistique, psychologie, sciences de l'éducation, traitement automatique du langage, médecine, biologie, mécanique, électronique) et dans des sous-genres variés (articles et communications écrites, thèses de doctorat, mémoires d'habilitation à diriger des recherches).

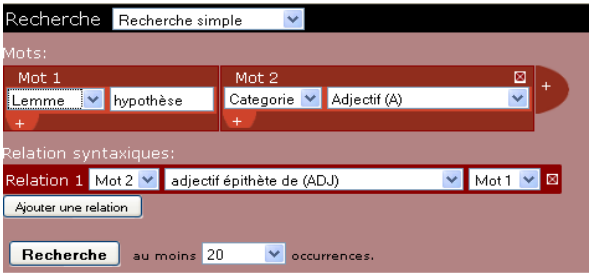
Le corpus a été annoté structurellement, en suivant les recommandations de la TEI Lite, et analysé syntaxiquement à l'aide de l'analyseur de dépendances Syntex, développé par Didier Bourigault (2007).

2.2 Les modes d'accès aux textes

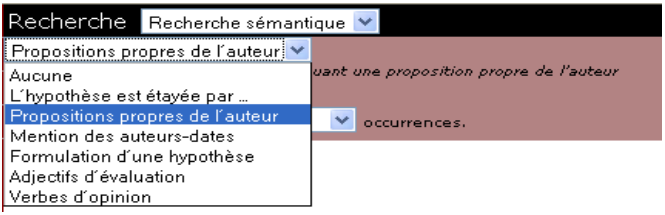
[1] Une large partie de ce corpus sera disponible pour la communauté de recherche.

Une fois le corpus sélectionné selon les disciplines, les genres textuels et les parties textuelles désirés, l'utilisateur peut accéder au contenu du texte par trois types de recherche :

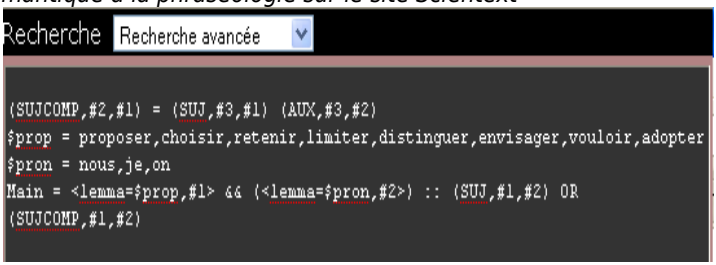
- **Un mode simple et guidé** avec des ascenseurs permet à l'utilisateur de sélectionner des formes, lemmes et/ou catégories, ainsi que les relations syntaxiques désirées. La figure ci-dessous permet ainsi d'extraire les occurrences de *hypothèse* avec un adjectif épithète.



- **Un mode sémantique** permet d'accéder à des occurrences en corpus, à partir de grammaires prédéfinies. Les grammaires sont définies à l'aide d'un langage de requête, ConcQuest, défini par Olivier Kraif, et étendu par nous.



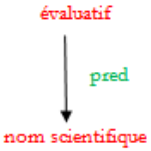
- **Un mode complexe** permet d'accéder à des occurrences en corpus, à partir de grammaires, utilisant les dépendances syntaxiques, les relations linéaires et des variables.



3. Le mode sémantique

Le mode sémantique suit des schémas sémantico-rhétoriques, qui sont ensuite traduits dans le langage de requête, à l'aide de variables et de dépendances syntaxiques.

A titre d'exemple, voici le schéma simple utilisé pour l'évaluation adjectivale :



On lui fera correspondre une grammaire utilisant pour chaque notion (en rouge) un ensemble de lexèmes, alors que la relation pred (en vert) sera traduite par la relation syntaxique épithète ou attribut, comme indiqué ci-dessous.

```
//TITRE: Adjectifs d'évaluation

//INFO: Les adjectifs d'évaluation qui portent sur les noms scientifiques

(ATTRIB,#2,#1) = (SUJ,#3,#1) (ATTS,#3,#2) ;
$eval=acceptable,adéquat,aisé,ambitieux,approximatif,bon,clair,classique,cohérent,complexe,concis,confus,convaincant,correct,crucial,déterminant,difficile,discutable,encourageant,épineux,essentiel,excellent,faible,fin,flou,fondamental,important,innovant,intéressant,irréprochable,judicieux,majeur,mauvais,meilleur,important,pertinent,nouveau,original,passable,passionnant,performant,principal,prometteur,riche,rigoureux,satisfaisant,séduisant,sérieux,significatif,solide,souhaitable,stimulant,vague,valable
$theo=
analyse,approche,article,caractéristique,choix,communication,concept,contribution,critère,élément,étude,exemple,facteur,fonction,idée,méthode,modèle,notio
```

```
n,objectif,phénomène,problème,projet,proposition,qualité,question,réflexion,résultat,solution,test,théorie,travail

Main = <lemma=$eval,#1> && <lemma=$theo,#2> :: (ATTRIB,#1,#2) OR (ADJ,#1,#2);
```

Une fois la requête lancée, il sera possible d'obtenir un affichage des occurrences dans un concordancier, ainsi que des extractions statistiques.

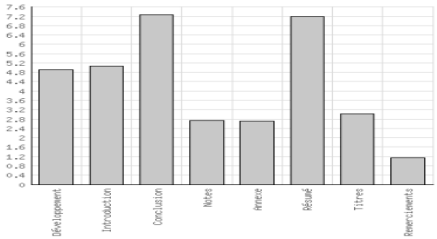
La figure suivante indique ainsi les occurrences les plus fréquentes, alors que la figure d'après indique les disciplines où ces expressions apparaissent proportionnellement les plus fréquentes.

Au total 2068 occurrences ont été trouvées.

Liste des lemmes

Lemme	occurrences
meilleur résultat	53
proposition principal	44
objectif principal	44
bon résultat	36
problème majeur	29
principal caractéristique	28
bon qualité	27
principal résultat	25
modèle classique	23
élément essentiel	22
meilleur qualité	21
problème complexe	21
analyse fin	21
facteur important	20
résultat intéressant	19
caractéristique principal	19
phénomène cohérent	19
résultat satisfaisant	18

Partie textuelle	Nombre absolu d'occurrences		Nombre de mots total		Nombre relatif d'occurrences
Développement	1789	/	3645711	=	4.91 ‰
Introduction	106	/	209266	=	5.07 ‰
Conclusion	62	/	85200	=	7.28 ‰
Notes	42	/	153355	=	2.74 ‰
Annexe	32	/	118198	=	2.71 ‰
Résumé	18	/	25061	=	7.18 ‰
Titres	17	/	56124	=	3.03 ‰
Remerciements	2	/	17551	=	1.14 ‰



4. Bibliographie

Bourigault Didier (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Mémoire d'Habilitation à Diriger les Recherches, Toulouse.

Kraif Olivier (2008). Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest, *Actes des 9ème Journées d'analyse statistique des données textuelles, JADT 2008*. Lyon: Presses universitaires de Lyon: 625-634.

Tutin Agnès (à paraître). Showing phraseology in context: an onomasiological access to lexico-grammatical patterns in corpora of French scientific writings, *Proceedings of eLexicography in the 21st century: new challenges, new applications, 22-24 october 2009, Louvain la Neuve.*

5. A propos des auteurs

Falaise Achille

Laboratoire GETALP, LIG

Post-doctorant en informatique

Thèmes de recherche : TAL, traduction automatique, écrits électroniques, IHMGETALP-LIG

385 rue de la Bibliothèque - BP 53

38041 Grenoble Cedex 9

achille.falaise@imag.fr

http://www-clips.imag.fr/geta/User/achille.falaise/

Tutin Agnès

LIDILEM, Université Grenoble3-Stendhal

Maître de conférence (HDR) en linguistique
Thèmes de recherche : linguistique de corpus, phraséologie, écrits scientifiques, TAL

UFR des Sciences du Langage
BP 25 - 38040 Grenoble cedex 9

agnes.tutin@u-grenoble3.fr

www.u-grenoble3.fr/tutin