Routines sémantico-rhétoriques de l'écrit scientifique

Agnès Tutin – Olivier Kraif – Sylvain Hatier
Laboratoire Lidilem – Université Grenoble Alpes

Journée phraséologique franco-polonaise - 14 nov. 2015

Introduction

- Contexte de l'étude
  - Dans la continuité du Projet Scientext et dans le cadre du projet Termith(2013-2016)
    - Financement ANR
    - Université de Nancy (ATILF), LINA (Nantes), LORIA (Nancy)
  - Etudier le lexique transdisciplinaire des écrits scientifiques : extraction, analyse et traitement
  - Utiliser le lexique scientifique transdisciplinaire (LST) pour faciliter le processus d'indexation dans les textes scientifiques des sciences humaines

Plan

- Qu'appelle-t-on phraséologie scientifique transdisciplinaire?
- Différents types d'expressions polylexicales
- Des motifs aux routines
- Extraction des routines à l'aide des arbres lexico-syntaxiques récurrents
- Perspectives

Qu'appelle-t-on phraséologie scientifique transdisciplinaire?

- **Phraséologie spécifique au genre des écrits scientifiques** et surreprésentée (voir Pecman 2007; Kosem 2010; Paquot 2012)
  - N'est pas une terminologie scientifique p.e. *entrée lexicographique*
  - N'est pas une terminonologie universitaire (Cf. Granger *et al.* 2013)
- **Porte sur**
  - **L'activité scientifique et l'évaluation** : *collecter des données, mener une expérimentation, résultats encourageants*
  - **Le raisonnement scientifique** : *nous arrivons à la conclusion, c'est pourquoi …*
  - **Le métadiscours et le métatexte :** *dans un premier temps, comme on l'a vu, au contraire*

- Les expressions polylexicales sont très fréquentes dans les textes scientifiques

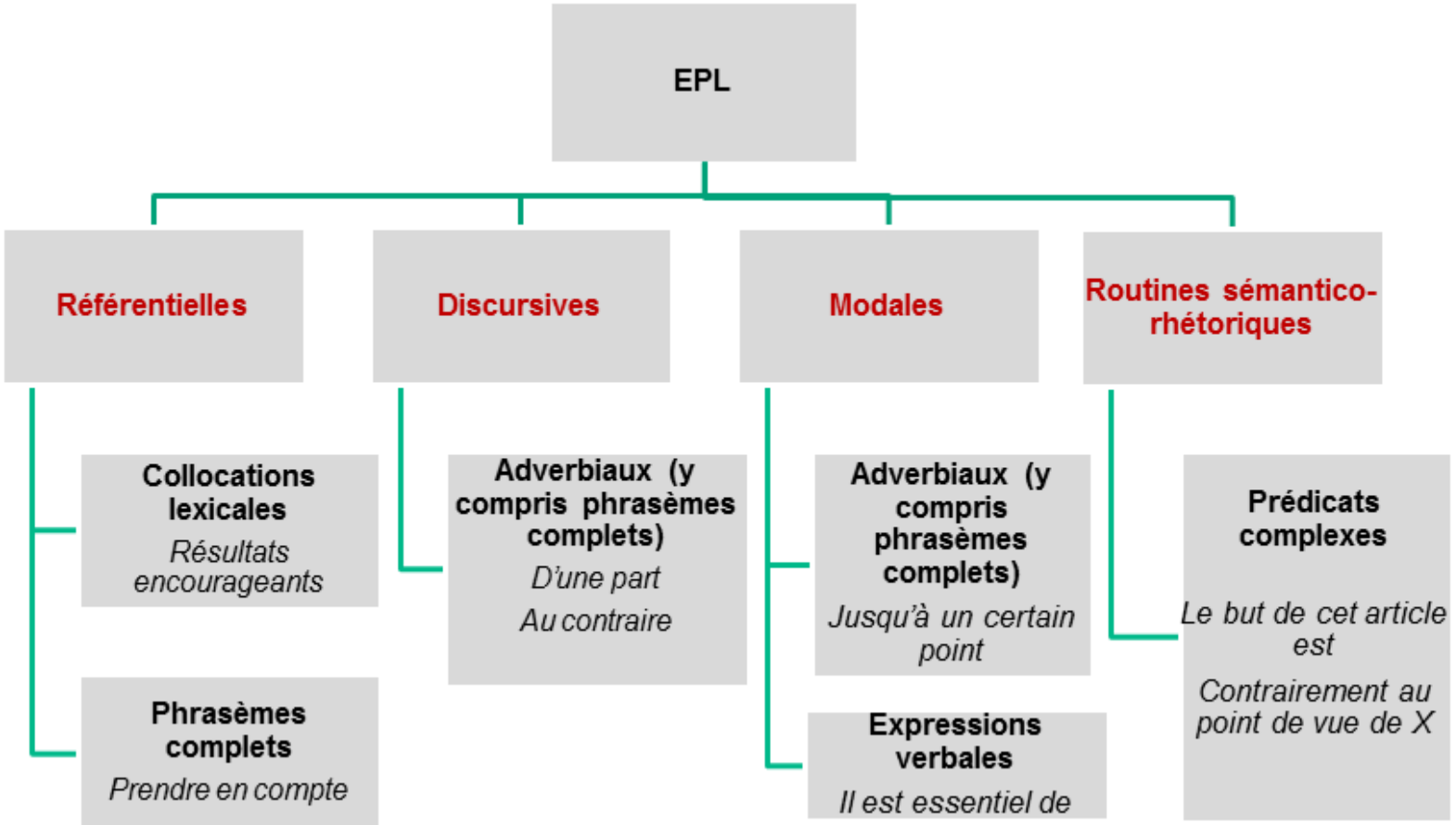
En premier lieu, nous souhaitons défendre la thèse selon laquelle les expressions polylexicales répondent à des régularités. Nous remettons en question la thèse anomaliste et proposons jusqu'à un certain point un point de vue analogiste.

LIDILEM

UNIVERSITÉ Stendhal GRENOBLE 3

Différents types d'expressions polylexicales

- Les écrits scientifiques comprennent tous les types d'expressions polylexicales, à l'exception des proverbes et pragmatèmes (p.e. *y'a pas de quoi, après vous!*)
- Une typologie fonctionnelle et structurale (inspirée de Granger & Paquot 2008; Burger 1998; Mel'čuk 2011)



Différents types de modélisations pour différents types d'EPL

Une typologie structurale (syntaxe et sémantique) est aussi nécessaire pour :

- La phraséologie basée sur corpus
- Le traitement lexicographique des EPL

Type d'EPL	Techniques d'extraction	Traitement lexicographique
Collocations référentielles	Mesures d'association (p.e. log likelihood ratio)	Traitement complexe : expressions compositionnelles (p.e. Fonctions Lexicales)
Phrasèmes complets	N-grammes lexico-syntaxiques (qui utilisent des dépendances syntaxiques)	Traitement complexe pour les verbes
EPL discursives	N-grammes lexicaux	EPL équivalents à des mots simples
EPL modales		
Routines sémantico-rhétoriques	Arbres lexico-syntaxiques récurrents	Représentations complexes (à l'aide de cadres)

Des motifs aux routines

- 3-4 diapos

Extraction des routines à l'aide des arbres lexico-syntaxiques récurrents

- Corpus : 5 M de mots, 10 domaines :

anthropologie (ANT), économie (ECO), histoire (HIS), géographie (GEO), linguistique (LIN), psychologie (PSY), sciences de l'éducation (SCE), sciences de l'information (SCI), sciences politiques (SCP), sociologie (SOC)

- Annoté en dépendances avec XIP
- Étude autour des verbes de constat :

constater, voir, noter, remarquer, observer

LIDILEM

UNIVERSITÉ Stendhal GRENOBLE 3

Extraction des routines à l'aide des arbres lexico-syntaxiques récurrents

- La méthode des ALR donne des résultats correspondant à des types variés, avec de nombreuses routines :

  < *force est de constater que* >, < *comme nous l'avons vu* >

  < *il est intéressant de noter* >, < *nous allons le voir* >

- Des variations importantes en surface ne font pas obstacle à l'identification :

SCP-12572 : Sans doute **faut -il y voir** un effet de la dissociation

SOC-12253 : **Il faut** à nouveau **y voir**, mais pas seulement, cette nécessaire adaptation

ANT- 2214 : [...] **il** n' **y faut voir** aucune trace d' une quelconque « modernité »

→ intérêt du recours aux relations de dépendance

Extraction des routines à l'aide des arbres lexico-syntaxiques récurrents

- La définition d'une classe $CONSTAT permet de grouper les observations et d'identifier plus de configurations intéressantes :

< On $CONSTAT en effet>

*ECO-8516 : **On remarque en effet** que, pour les hommes comme pour les femmes(...)*

*SCI-10116 : **En effet**, **on voit** bien que la relation de proximité instituée par le talk-show*

< On peut $CONSTAT ainsi>

*HIS-246 : **On peut ainsi voir** une représentation de Luther portant un crucifix et la Bible avec , derrière lui , un cygne représentant Hus lisant également la Bible .*

*SCE-5837 : **Ainsi** , par exemple , a **-t-on pu voir** des listes de mots auxquels on pouvait se référer le jour suivant .*

Extraction des routines à l'aide des arbres lexico-syntaxiques récurrents

- Il apparaît que certains ALR impliquent l'ensemble de la classe $CONSTAT
  - *<il est intéressant de $CONSTAT> : noter/constater/observer/remarquer/voir*
  - *<comme on pouvoir le $CONSTAT> : voir/constater/remarquer/noter*
  - *<il est à $CONSTAT> : noter/remarquer/observer/voir*

Extraction des routines à l'aide des arbres lexico-syntaxiques récurrents

- Un même ALR permet d'identifier à la fois des réalisations canoniques et des variantes plus rares. P.ex., pour <il est à $CONSTAT>, on trouve 18 occurrences de *il est à noter* avec des exceptions intéressantes *:*

*HIST-3862 - **Il est à remarquer** que le mouvement qui porte les nobles à se hisser dans la direction du capitalisme agro - industriel*

*SOC-2183 : Si l'on se rapporte à l' étude statistique menée à partir des questionnaires envoyés aux artistes , **il est à observer** que , pour la diffusion sur les scènes nationales entre 1998 et 2000*

Extraction des routines à l'aide des arbres lexico-syntaxiques récurrents

- Expansion des ALR : on fait l'hypothèse qu'à chaque position dans l'arbre correspond un paradigme sous-jacent :

<il être intéressant de $CONSTAT>

  - <il être intéressant de V> : {*noter/constater/observer/remarquer/comparer/relever/voir*}

<Il est frappant de $CONSTAT>

  - <Il est ADJ de $CONSTAT> : {*frappant/important/intéressant/possible*}

Extraction des routines à l'aide des arbres lexico-syntaxiques récurrents

<on $CONSTAT un différence>

- <on $CONSTAT un N> :
  {différence/effet/tendance}

<comme on $LE avoir $CONSTAT>

- <comme PRO $LE avoir $CONSTAT> :
  {on/nous}

Perspectives

- Construction de méta-grammaires permettant de caractériser, en une requête, les schémas de transformation productifs (avec les % d'occurrence) et les paradigmes liés :
- p.ex. : observer+différence :
  - % passif : une différence a été observée
  - % passif réduit : la différence observée
  - % insertion ADJ : observer une différence importante/notable/marquante
  - % insertion ADV : rarement/souvent observé une différence
  - etc.