Expressions polylexicales dans le discours scientifique : une base de données lexicales basée sur corpus

Agnès Tutin
LIDILEM, UFR Sciences du langage
Université de Grenoble Alpes
Achille Falaise
ICAR, ENS de Lyon

EUROPHRAS
10-12 septembre 2014, Paris

Cadre du projet

- Projet en cours sur le lexique scientifique transdisciplinaire (mots simples et expressions polylexicales) basé sur le corpus Scientext, un corpus d'écrits scientifiques du français et de l'anglais
- Base de données d'expressions phraséologiques basée sur corpus : « dictionary cum corpus »
  - Application lexicographique et pédagogique du projet Scientext
- Deux applications visées :
  - Objectifs pédagogiques (aide à la rédaction pour le français universitaire)
  - Lexique scientifique transdisciplinaire pour une application de traitement automatique du langage (projet Termith)

Utilisation du lexique scientifique transdisciplinaire pour l'indexation automatique des textes scientifiques

- Techniques d'indexation utilisant le lexique scientifique transdisciplinaire comme lexique d'exclusion (projet Termith)
  - Extraction of the main terms of the text
    - Un mot (ici *sujet*) ne peut pas être un candidat terme s'il est inclus dans une collocation transdisciplinaire

… différents champs théoriques **abordent** ce **sujet**.

'aborder un sujet' = collocation scientifique

    - Autrement, il peut être un candidat terme

Il s'agit fréquemment du **sujet** de la **proposition** principale…

'sujet de la proposition' = pas phraséologique

LIDILEM
UNIVERSITE Stendhal GRENOBLE 3

Plan

- Qu'appelle-t-on phraséologie scientifique transdisciplinaire?
- Différents types d'expressions polylexicales
- Méthodologie
- Structure de la base de données : l'exemple des expressions polylexicales à fonction modale

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*

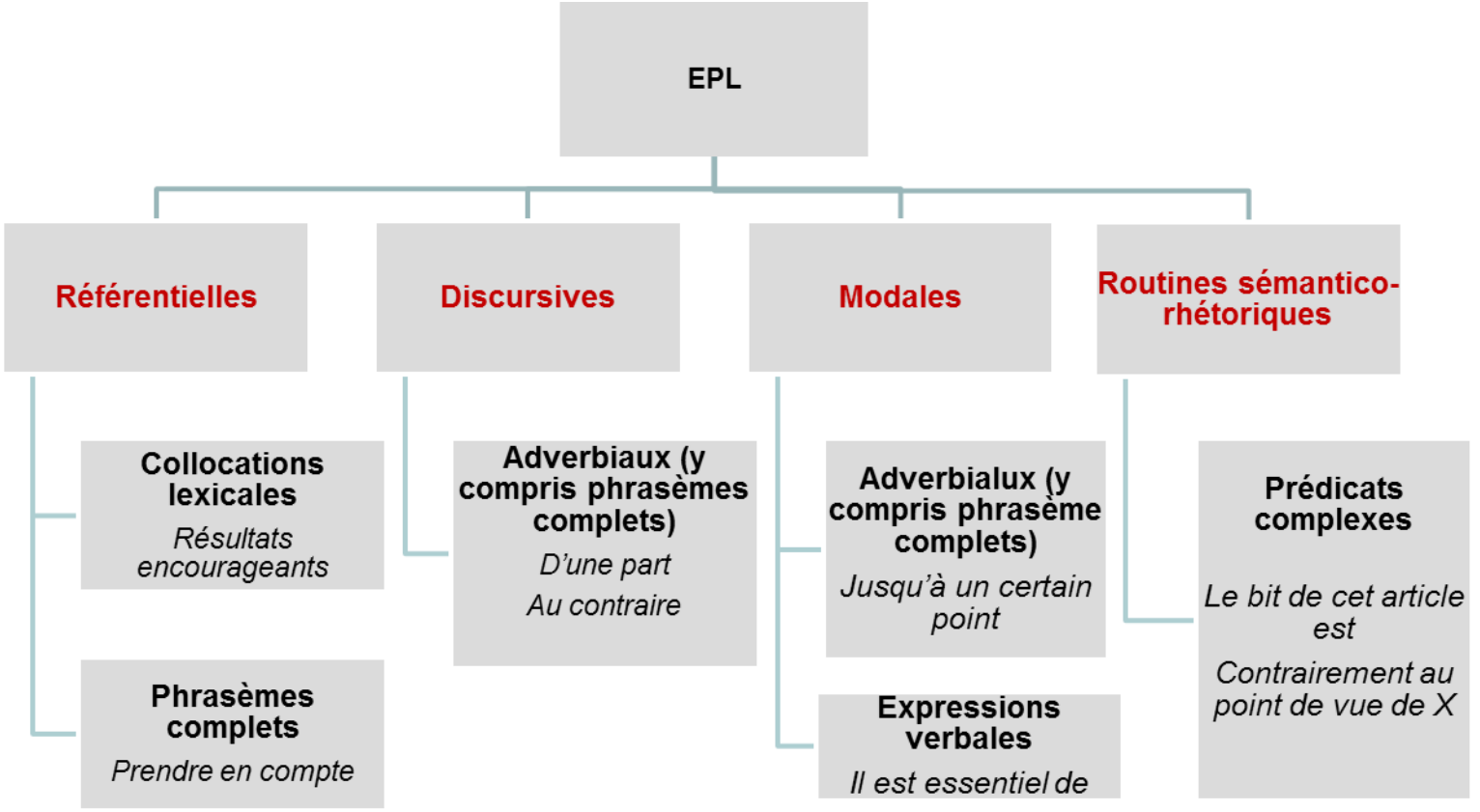
Exemple

- 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.

Les écrits scientifiques comprennent tous les types d'expressions polylexicales

- A 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 EPL modales	N-grammes lexicaux	EPL équivalents à des mots simples
Routines sémantico-rhétoriques	N-grammes syntaxiques et sémantiques ?	Représentations complexes (à l'aide de cadres)

Exemples de n-grammes lexico-syntaxiques

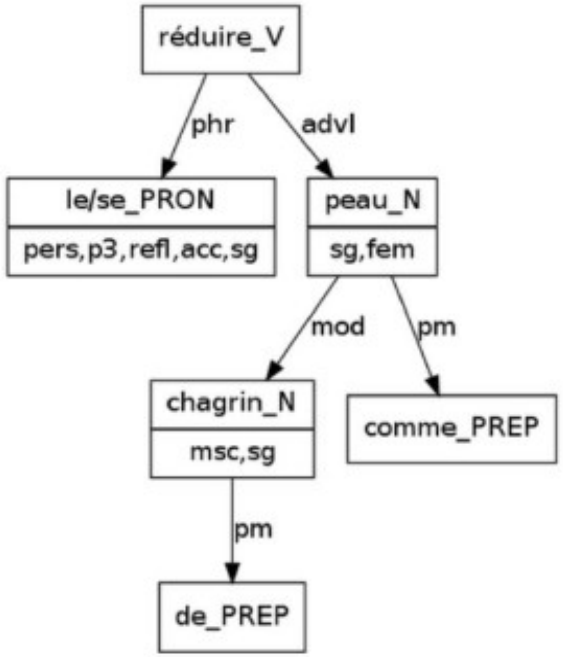
- Structures lexico-syntaxiques récurrentes (les mots ne sont pas nécessairement contigus) (utilisant une mesure PMI et des relations syntaxiques) (Corman 2012)
- Le/se réduire à une peau de chagrin :

DÉPENDANCES ET TRAITS PRIVILÉGIÉS

```
réduire_V -> le/se_PRON{pers,p3,refl,acc,sg} ^^ réduire_V ->
peau_N{sg,fem,prep:comme_PREP,det:NO_DET} ->
chagrin_N{prep:de_PREP,msc,sg,det:NO_DET}
```

OCCURRENCES : 69

MOTS PLEINS : 4



2. Méthodologie

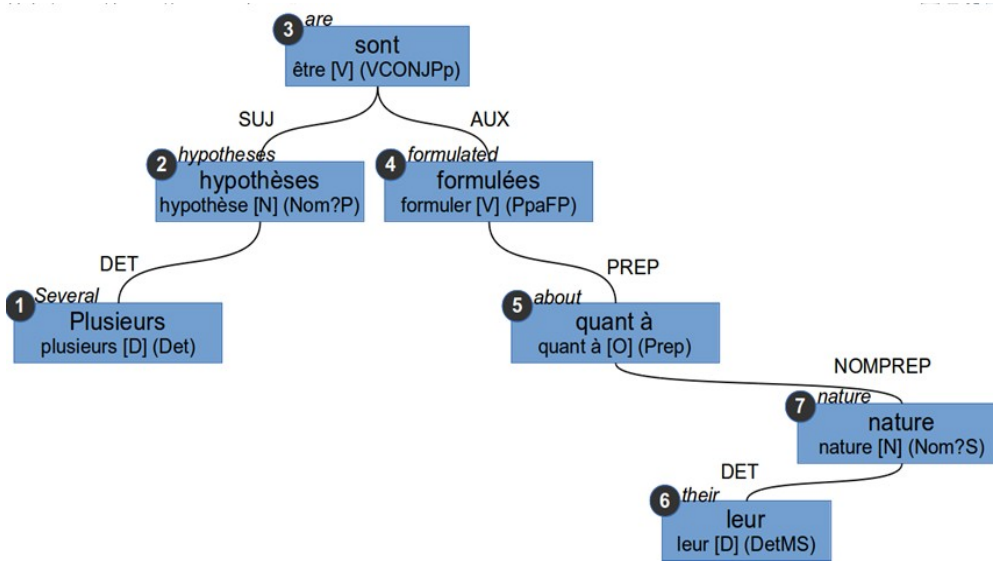
- Notre base de données phraséologique est basée sur plusieurs principes simples :
  - **Ressource basée sur corpus** :
    - Toutes les ressources sont basée sur un corpus en-ligne d'écrits scientifique, Scientext, librement accessible sur Internet.
  - **Accès onomasiologique et sémasiologique** (→ Pecman 2007; Granger & Paquot 2010)
    - Accès onomasiologique par le biais d'étiquettes sémantiques, et accès sémasiologique par les entrées lexicales. Informations : sémantique, syntaxe, fonctionnement discursif, fréquence, synonymes.
  - **Projet "Dictionary-cum-corpus"** (Hartmann 2005; Granger & Paquot 2010)
    - Les EPL sont associées à des exemples triés en ligne, avec des contextes larges.
    - Une information sur l'usage de l'expression est proposée

Etape 1 : Extraction

- Extraction à partir d'un corpus annoté syntaxiquement (Syntex, Bourigault 2007), comportant des annotations sur les parties textuelles (introduction, conclusion, notes …)

– Utile car les composantes des EPL ne sont pas nécessairement contiguës et pratique pour traiter les alternances syntaxiques comme le passif (Seretan 2011)

– Plusieurs techniques selon les EPLs à repérer : n-grammes, mesures d'association des collocations (Kilgarriff *et al.* 2004; Seretan 2011 ; Diwersy & Kraif 2012), n-grammes lexico-syntaxiques(Corman 2012).



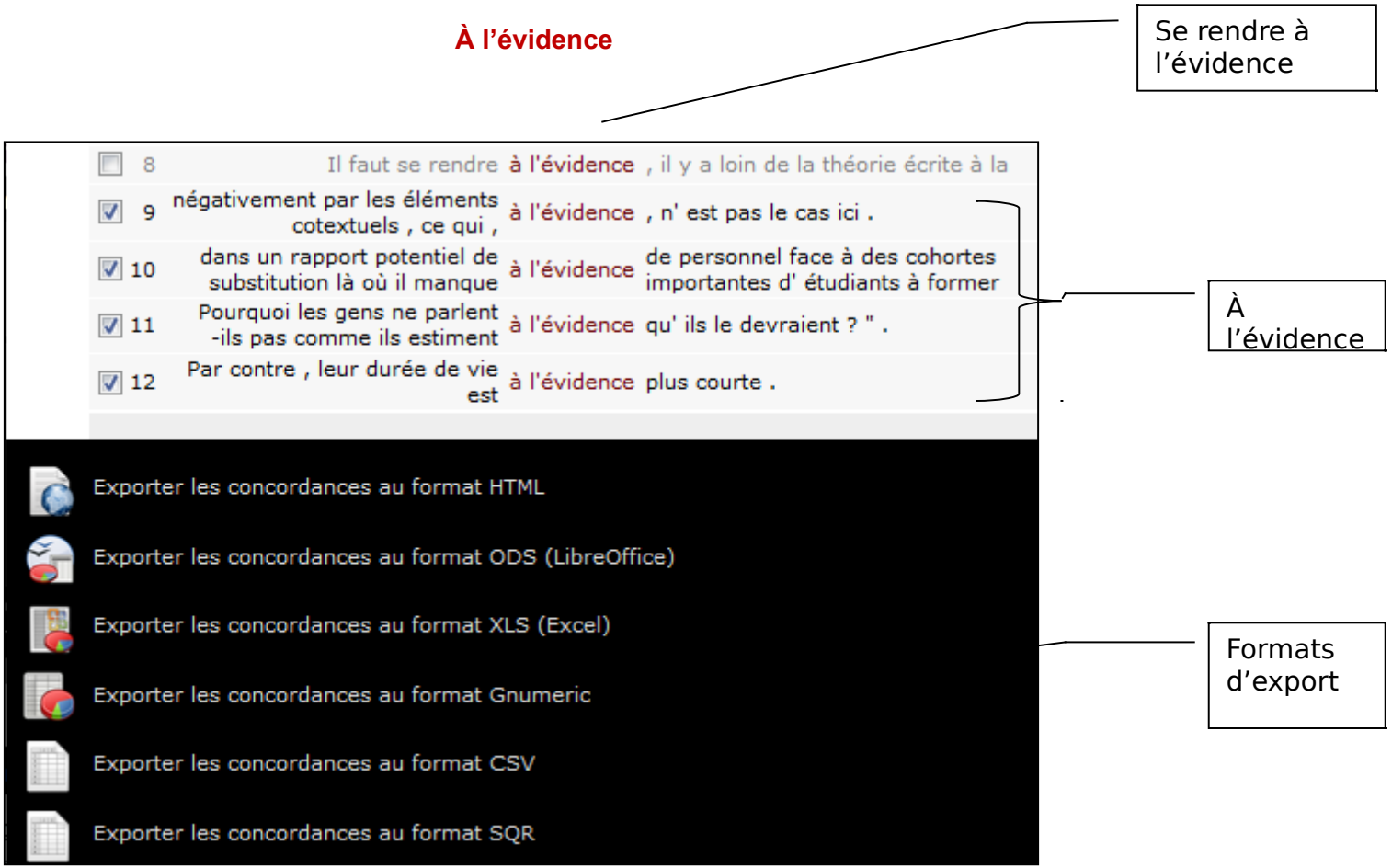
Etape 1 : Extraction

- Corpus Scientext : 234 textes (1997-2008), 5 million de mots (accessible sur le web; license CC) (8 millions de mots avec le corpus interne), avec des annotations syntaxiques et sur les structures textuelles.
- Corpus pas parfaitement équilibré (pour des raisons de droits). La demande principale pour l'aide à la rédaction est en sciences sociales et humaines

	Articles & communications	Thèses	HDRs
SHS	**154**	**32**	**8**
Linguistique	66	8	4
Sciences de l'éducation	63	8	2
TAK	8	8	1
Psychologie	17	8	1
Sciences naturelles	**21**	**10**	**0**
Biologie	6	7	0
Médecine	15	3	0
Sciences appliquées	**0**	**8**	**1**
Electronique	0	4	0
Mécanique	0	4	1
TOTAL	**175**	**50**	**9**

Etapes 2 et 3 : Traitement linguistique des EPL extraites et sélection des exemples en ligne

- Traitement lexicographique des EPL extraites
  - Information syntaxique, sémantique, discursive, de fréquence
- Sélection des exemples pertinents



L'exemple des EPL à fonction modale

- Plusieurs bases en développement :
  - Selon le type d'EPL
    - Base des EPL à fonction discursive (Tran)
  - Structure simple
    - Plus d'attention est prêtée aux exemples en corpus

- EPL à fonction modale :
  - Expriment le point de vue, l'attitude au sens large, et sont centrales dans l'écrit scientifique (Hyland 2008)
  - Simples à extraire. La plupart d'entre elles sont des adverbes (Extraction à l'aide n-grammes)

Champs de l'entrée lexicale

1. **Entrée lexicale et variantes**

- Par exemple, *de notre point de vue* a des variantes de personne *de mon point de vue*.

2. **Types sémantiques** (accès onomasiologique)

- certitude', 'évidence', 'opinion', 'opinion positive', 'opinion négative', 'approximation', 'necessité', 'probabilité ', etc.

➢ Assez complexe, du fait de la polysémie

3. **Partie du discours de l'EPL**

- p.e. : *à l'évidence* : Adverbe

4. **Parties du discours des composants**

1. *P.e. :* à *:* Prep *l' :* Det *évidence :* Nom

➢ Utile pour l'extraction à partir des corpus

LIDILEM UNIVERSITE Stendhal GRENOBLE 3

Champs (suite)

5. **Frequence**.

- Très utile pour les applications pédagogiques. Valeurs approximatives (***: fréquent, **: assez fréquent, *: pas si fréquent). Une valeur précise le nombre d'occurrences valides trouvées dans le corpus Scientext.

6. **Synonymes, de préférence des expressions polylexicales**

7. **Définitions courtes**

8. **Lien à des exemples vérifiés**

Champs (suite)

9. Distribution syntaxique des EPLs

•Un problème complexe pour les étudiants non-natifs (e.g. Osborne 2008 en Anglais.

1	C. Fuchs insiste par ailleurs	à juste titre	sur le caractère formateur et l' intérêt didactique de la
2	faire retomber dans ses " motivations extérieures " que dénonce	à juste titre	B. Lahire .
3	la lecture " , parue en novembre 2000 , remarque d'ailleurs	à juste titre	le peu d' études qui ont été consacrées au niveau morphématique dans
4	57) souligne	à juste titre	que la restauration de la dimension argumentative dans le travail

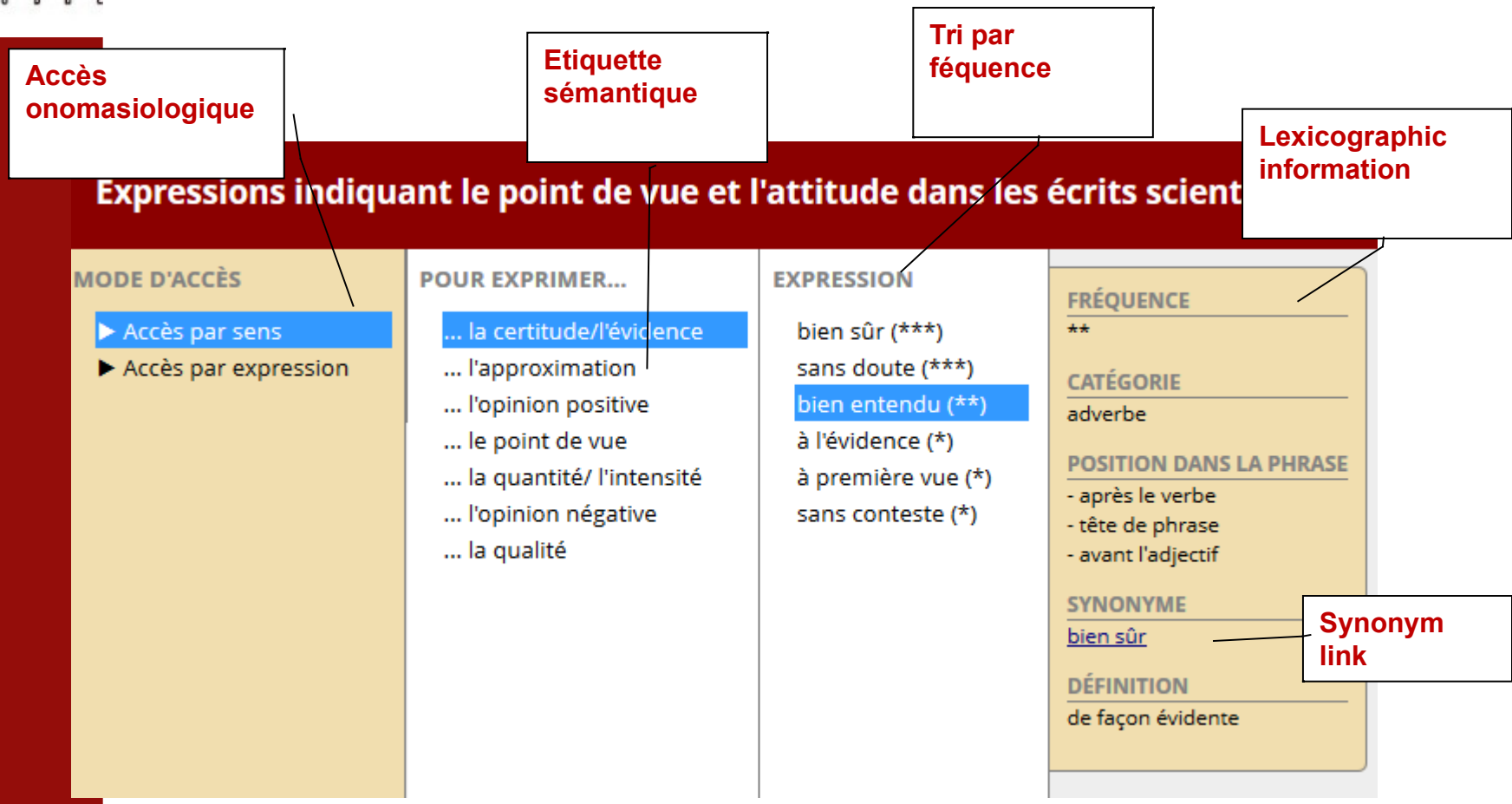
À juste titre = Après le verbe

Position initiale dans la phrase

postverbal

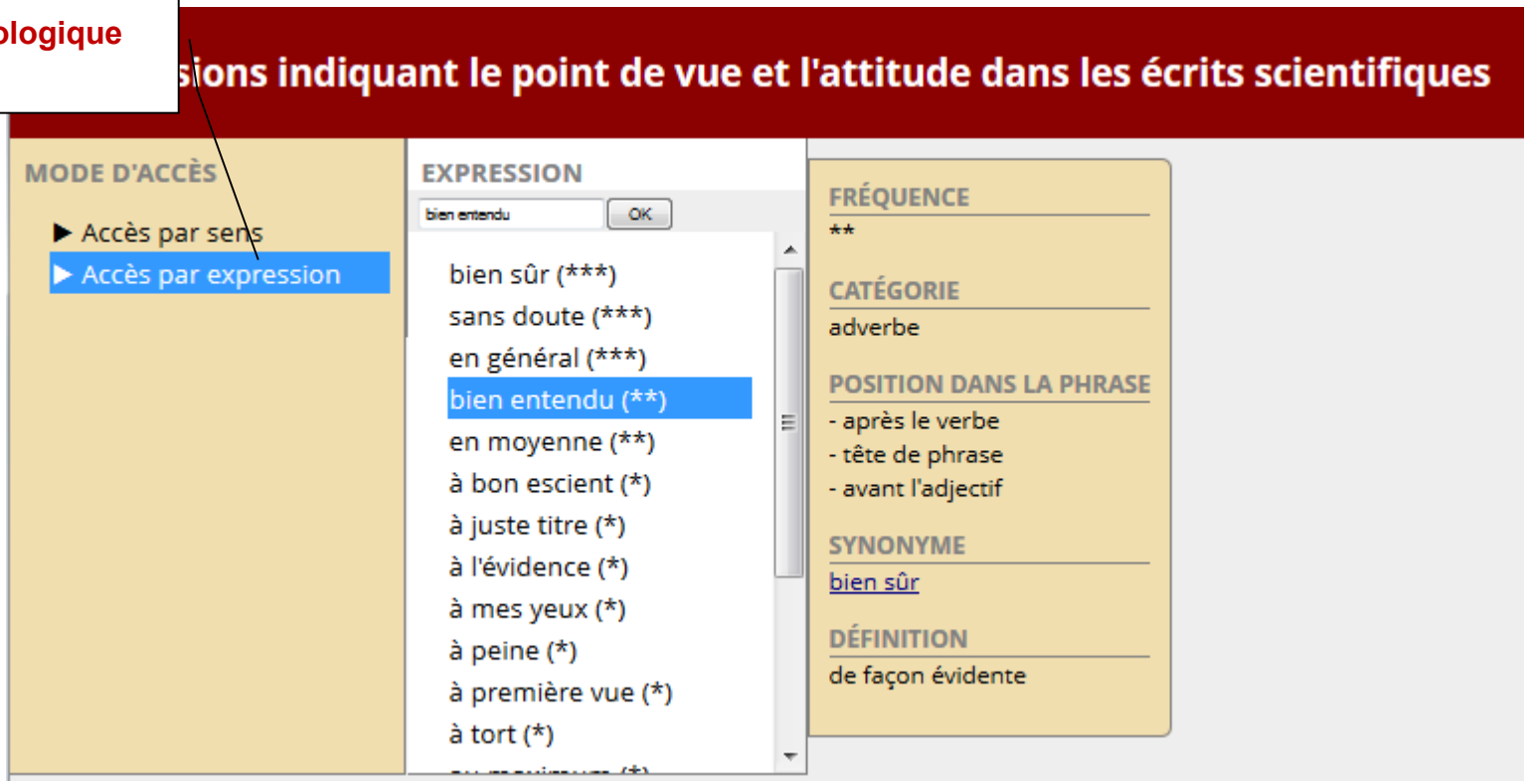
6	[45] Bien entendu	,	cette affirmation est à prendre avec prudence et surtout
7	Je parle	bien entendu	des enfants .
8		Bien entendu	, le contexte des LP est tout différent , et
9	Enfin ,	bien entendu	, l' expression renvoie aussi implicitement à la sociologie compréhensive
10	Il s' agit là	bien entendu	d' une vision trop simple , dans la mesure où d' autres espaces

Accès onomasiologique à la base de données



Accès sémasiologique à la base de données

Accès sémasiologique



Information d'usage
Fréquence relative dans les textes scientifiques et en langue générale

Bien entendu

Fréquence relative dans les écrits scientifiques. Calculée avec les exemples sélectionnés

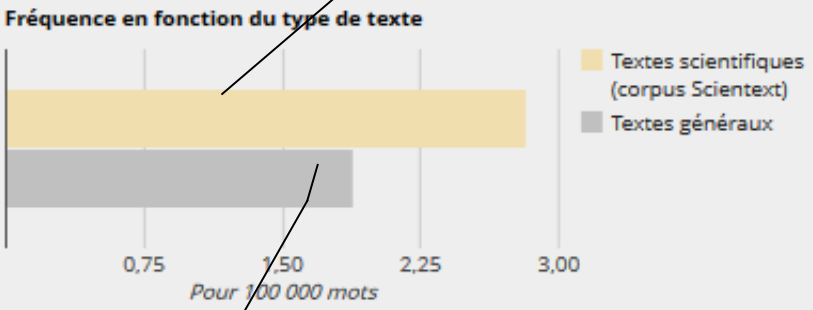
FRÉQUENCE
**

CATÉGORIE
adverbe

POSITION DANS LA PHRASE
- après le verbe
- tête de phrase
- avant l'adjectif

SYNONYME
bien sûr

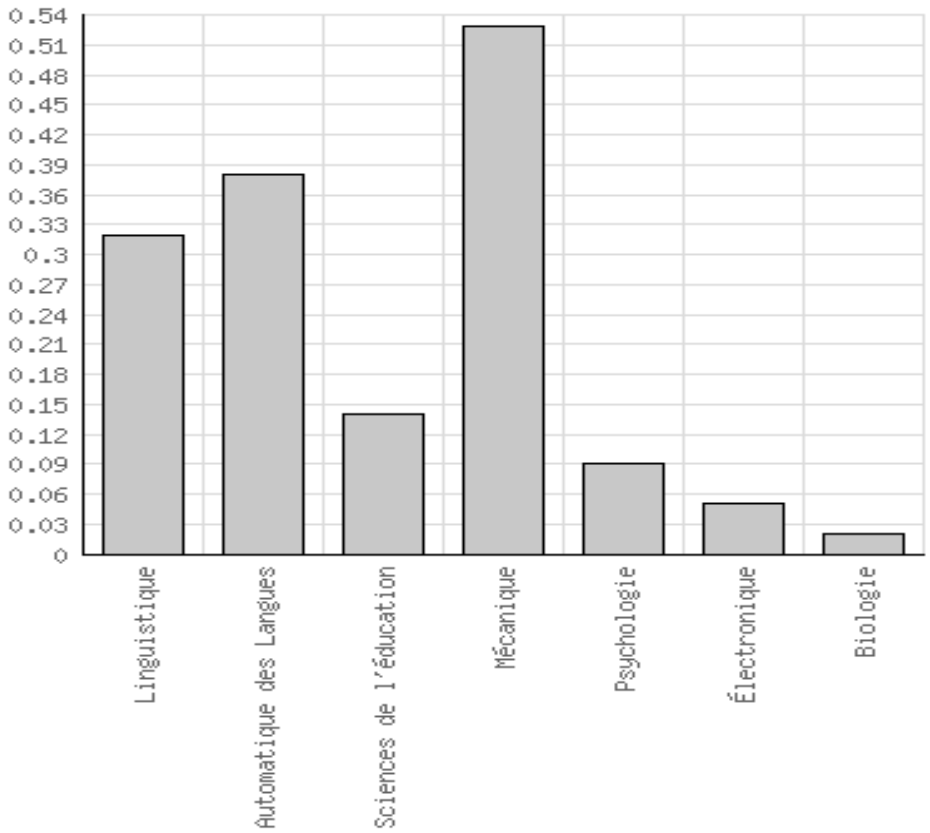
DÉFINITION
de façon évidente



Fréquence relative dans la langue générale. Calculée sur un corpus de 30 millions de mots (presse et romans)

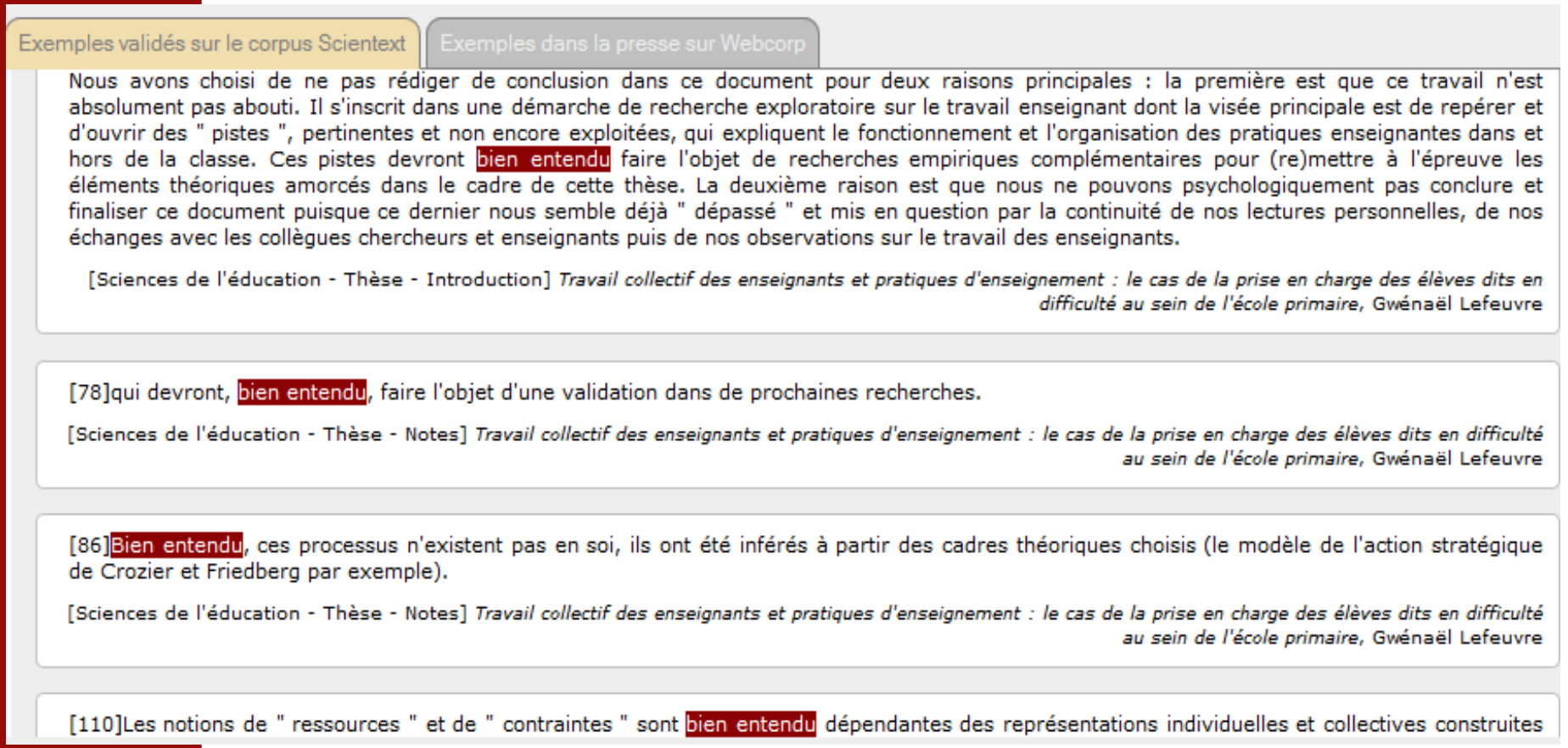
Distribution de l'expression selon la discipline, partie textuelle, type de texte (en cours)

- Distribution de *bien entendu* selon la discipline (fréquence relative)



Concordances de textes scientifiques sélectionnées en ligne (Corpus Scientext)

- **Concordances larges (Les concordances KWIC ne conviennent pas pour l'aide à la rédaction)**
- **Contextes larges (→ 300 mots) disponibles**



Exemples validés sur le corpus Scientext | Exemples dans la presse sur Webcorp

Nous avons choisi de ne pas rédiger de conclusion dans ce document pour deux raisons principales : la première est que ce travail n'est absolument pas abouti. Il s'inscrit dans une démarche de recherche exploratoire sur le travail enseignant dont la visée principale est de repérer et d'ouvrir des " pistes ", pertinentes et non encore exploitées, qui expliquent le fonctionnement et l'organisation des pratiques enseignantes dans et hors de la classe. Ces pistes devront bien entendu faire l'objet de recherches empiriques complémentaires pour (re)mettre à l'épreuve les éléments théoriques amorcés dans le cadre de cette thèse. La deuxième raison est que nous ne pouvons psychologiquement pas conclure et finaliser ce document puisque ce dernier nous semble déjà " dépassé " et mis en question par la continuité de nos lectures personnelles, de nos échanges avec les collègues chercheurs et enseignants puis de nos observations sur le travail des enseignants.

[Sciences de l'éducation - Thèse - Introduction] *Travail collectif des enseignants et pratiques d'enseignement : le cas de la prise en charge des élèves dits en difficulté au sein de l'école primaire*, Gwénaël Lefeuvre

[78]qui devront, bien entendu, faire l'objet d'une validation dans de prochaines recherches.

[Sciences de l'éducation - Thèse - Notes] *Travail collectif des enseignants et pratiques d'enseignement : le cas de la prise en charge des élèves dits en difficulté au sein de l'école primaire*, Gwénaël Lefeuvre

[86]Bien entendu, ces processus n'existent pas en soi, ils ont été inférés à partir des cadres théoriques choisis (le modèle de l'action stratégique de Crozier et Friedberg par exemple).

[Sciences de l'éducation - Thèse - Notes] *Travail collectif des enseignants et pratiques d'enseignement : le cas de la prise en charge des élèves dits en difficulté au sein de l'école primaire*, Gwénaël Lefeuvre

[110]Les notions de " ressources " et de " contraintes " sont bien entendu dépendantes des représentations individuelles et collectives construites

Concordances de la langue générale sélectionnées (Presse, Webcorp)

Exemples validés sur le corpus Scientext | Exemples dans la presse sur Webcorp

1) http://www.lefigaro.fr/musique/2013/10/11/03006-20131011ARTFIG00491-nile-rodgers-a-bien-failli-perdre-sa-guitare-fetiche.php
Text, Wordlist, text/html, UTF8 (HTML source), 2013-01-01 (URL)

1: en empruntant le métro. Sa fidèle guitare est bien entendu du voyage, bien rangée dans son étui. Déconcentré

2) http://www.lefigaro.fr/culture/2013/10/02/03004-20131002ARTFIG00440--comment-ca-va-bien-l-art-de-savoir-faire-les-noeuds.php
Text, Wordlist, text/html, UTF8 (HTML source), 2013-01-01 (URL)

2: ces magazines féminins. Une rubrique «mode», bien entendu, tenue par la charmante Caroline Baly, qui nous

3) http://halleyjc.blog.lemonde.fr/2008/09/29/la-banane-notre-seul-medicament/
Text, Wordlist, text/html, UTF8 (Content-type), 2008-09-29 (URL)

3: BANANE, notre seul médicament ! après le Zouk ! Bien entendu ! Publié le 29 septembre 2008 par halleyjc
4: BANANE, notre seul médicament ! après le Zouk ! Bien entendu ! halleyjc dit : 30 septembre 2008 à 09:19

LIDILEM UNIVERSITE Stendhal GRENOBLE 3

Conclusion et travaux futurs

- **Une application lexicographique simple d'un corpus spécifique pour des besoins spécifiques**
  - **Une base de données lexicographiques avec des exemples basés sur corpus**
  - **Plus un corpus avec un accès lexical qu'un projet "dictionary-cim-corpus"**
- **Prototype de dictionnaire bientôt en ligne sur le site Scientext : http://scientext.msh-alpes.fr**
- **Conçu principalement comme aide à la rédaction mais doit aussi être évalué avec des apprenants du français langue étrangère**
- **Développements futurs :**
  - **La base de données doit être complétée avec d'autres types d'expressions**
  - **Quelques procédures devraient pouvoir être automatisées (p.e; position de l'EPL)**