L'annotation sémantique pour l'enseignement/apprentissage des marqueurs polylexicaux à fonction métadiscursive à l'appui du corpus

Thi Thu Hoai TRAN

Sylvain HATIER

LIDILEM, Université Grenoble Alpes

Thi-Thu-Hoai.Tran@u-grenoble3.fr

Sylvain.Hatier@u-grenoble3.fr

8es Journées Internationales de Linguistique de Corpus

2-4 septembre 2015, Orléans

PLAN

- Objectifs de recherche
- Cadre de recherche
- Méthodologie de recherche
- Typologie des séquences lexicalisées à fonction discursive
- Annotation sémantique des marqueurs discursifs pour des applications didactiques

OBJECTIFS DE RECHERCHE

Objectifs linguistiques

Relever les propriétés syntaxiques et sémantiques des séquences lexicalisées à fonction discursive

Objectifs didactiques

- Etablir des ressources lexicales au service de l'enseignement/apprentissage de ces unités lexicales
- Aider à la rédaction scientifique

❖Comment une annotation sémantique peut-elle être bénéfique à l'enseignement/apprentissage des MD?

❖Quelles sont les stratégies d'enseignement pour atteindre l'objectif?

CADRE DE RECHERCHE

- Analyse syntaxique et sémantique du Lexique Scientifique Transdisciplinaire et des MD membre du LST

DÉFINITION LST

- Définition Tutin (2007a) :
  - Lexique propre aux écrits scientifiques
  - Est un lexique de genre
  - Renvoie au discours sur les objets et procédures scientifiques
  - Abstrait et largement transdisciplinaire
  - Transversal et spécifique

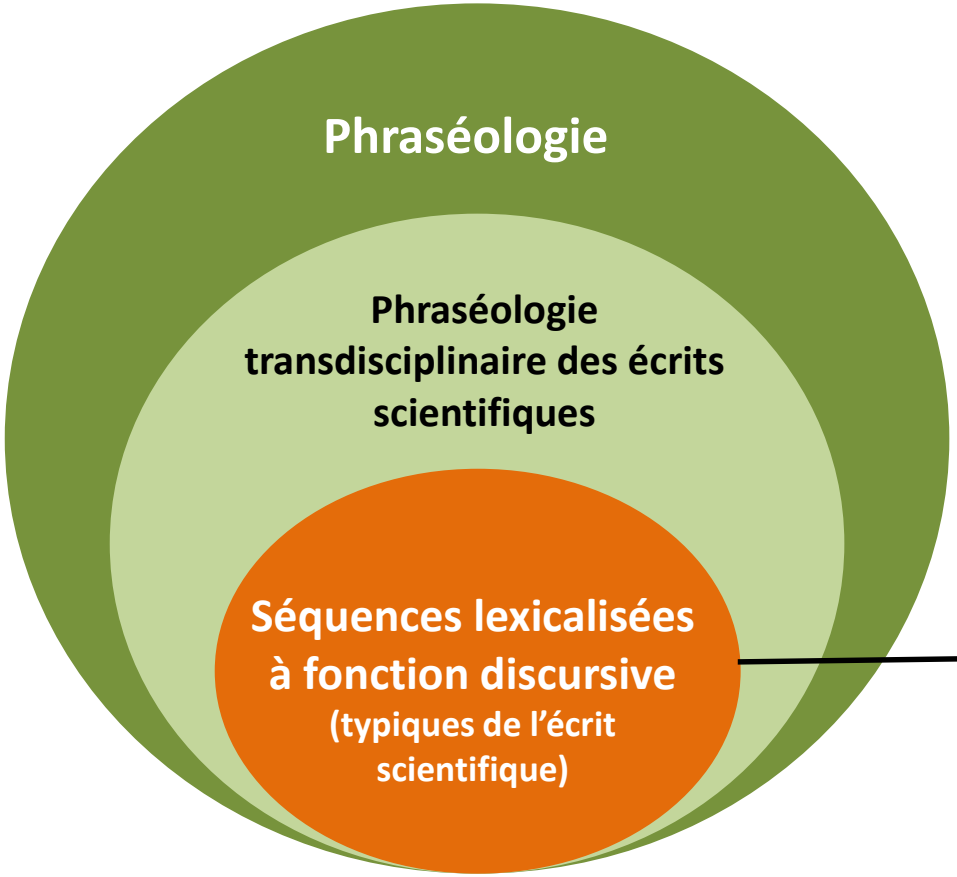
DÉFINITION LST

- Exprime/décrit les procédures scientifiques, le raisonnement, la filiation scientifique, le positionnement, le métatexte …
- Fréquemment en cooccurrence avec la terminologie
- Exemples :
  - *Nous [analyserons]$_{LST}$ uniquement les [textes]$_{T}$ des [bilingues]$_{T}$ dans cet [article]$_{LST}$*
  - *On [s'intéresse]$_{LST}$ dans cet [article]$_{LST}$ aux [travaux]$_{LST}$ sur la [morphologie]$_{T}$ des [adjectifs]$_{T}$*

CLASSIFICATION SÉMANTIQUE

- Extraction des unités lexicales répondant aux critères de spécification et de répartition
- Définition de classe et sous-classe sémantique structurant le LST :
  - Étiquette et définition
    - Ex : pour la classe *communication_contenu (affirmation, formulation, proposition, terme, commentaire, observation)*
      - *« contenu dans un processus de communication »*
  - Via le Lexicoscope, identification des relations lexico-syntaxiques communes à une classe pour élaborer des tests d'appartenance aux classes
    - *Par/dans ce N, l'auteur précise*
- Homogénéisation des classes à travers les catégories :
    - {*espace*} *: ailleurs, se situer, proche, lieu*

SÉQUENCES LEXICALISÉES À FONCTION DISCURSIVE (I)



✓ Différentes appellations : *marqueurs discursifs* (Dostie & Puch, 2007; Paillard & Vu, 2012), *particules énonciatives* (Fernandez, 1999), *marqueurs de glose* (Steuckardt & Niklas-Salminen, 2005).

✓ Rôle d'organisation textuelle et de structuration du discours.

Legallois, D., et Tutin, A. (2013)

SÉQUENCES LEXICALISÉES À FONCTION DISCURSIVE (II)

Nous formulons l'hypothèse que la démarche du lecteur, ou plus exactement d'une catégorie de lecteurs, peut d'une part, s'appuyer sur la présence de marques discursives et d'annotations dans le texte et, d'autre part, sur des connaissances qui exploitent celles-ci. De plus, ces connaissances sont susceptibles d'être modélisées sous une forme déclarative. En conséquence, nous proposons le langage SEXTANT. Par analogie avec les navigateurs du XVIIIè siècle qui ont parcouru le monde en s'orientant sur les mers avec un sextant. Pour exprimer ces connaissances. Autrement dit, nous considérons qu'il ne suffit pas de créer des liens mais qu'il est nécessaire d'expliciter l'opération de navigation et que, de plus, ce processus doit être mis en œuvre par un « expert » capable d'encoder ces connaissances. [tal-art-239-intro]

MÉTHODOLOGIE (I)

Linguistique de corpus

- ✓ Accéder aux données authentiques
- ✓ Repérer le fonctionnement syntaxique et sémantique des MD
- ✓ Critères statistiques
- ✓ Applications didactiques pour l'enseignement/apprentissage des séquences lexicalisées à fonction discursive

TAL

- ✓ Annotation semi-automatique des éléments du LST
- ✓ Identification des cooccurrents privilégiés des MD
- ✓ Repérage de patrons sémantico-syntaxique intégrant les MD

MÉTHODOLOGIE (II)

Annotation sémantique du LST

✓Injection dans le corpus des traits de classe et sous-classe sémantique

- En cas d'ambiguïté → union des traits

aussi : MD classe *{discursif}* et sous-classes {*explication}* et {*comparaison}*

```
1  <t c="PRON" l="le" f="MASC SG">le</t>
2  <t c="NOUN" l="monde" f="MASC SG">monde</t>
3  <t c="PREP" l="de" f="MASC SG DEF">du</t>
4  <t c="NOUN" l="voyageur" f="MASC SG P3">voyageur</t>
5  <t c="VERB" l="être" f="SG P3 PRE IND AUX">est</t>
6  <t c="ADV" l="aussi" f="ADV CS-discursif SCS-comparaison SCS-
   explication">aussi</t>
7  <t c="DET" l="un" f="MASC SG INDEF">un</t>
8  <t c="NOUN" l="laboratoire" f="MASC SG CS-personne SCS-collectif
   ">laboratoire</t>
```

MÉTHODOLOGIE (III)

Identification des cooccurrents syntaxiques privilégiés

- ✓ Application du rapport de vraisemblance (LLR) pour dégager les éléments en cooccurrence statistiquement significative
- ✓ Accès au profil combinatoire des MD

Marqueur discursif	Cooccurrents syntaxiques forts
en effet	*exister, constituer, apparaître*
autrement dit	*agir, devenir, étendre*
dans un premier temps	*aborder, proposer, présenter*
en ce sens	*employer, jouer, mener,*
dans ces conditions	*comprendre, concevoir, révéler*

MÉTHODOLOGIE (IV)

Identification des cooccurrents syntaxiques privilégiés

✓Parmi les cooccurrents, dégager les classes sémantiques le plus fréquemment en relation syntaxique

✓Ainsi, le MD de la sous-classe {énumération} *dans un premier temps* est souvent dépendant d'un verbe de la classe {examen} qui a pour membre : *explorer, examiner, aborder, étudier,* etc.

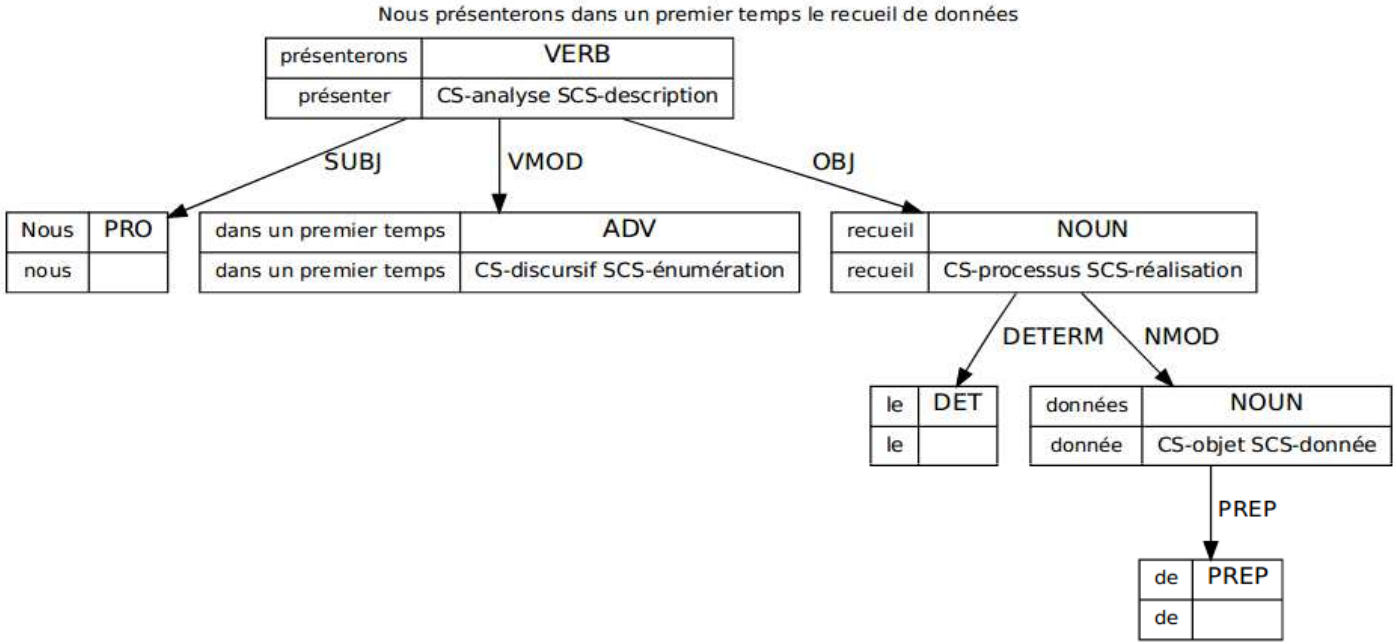
MÉTHODOLOGIE (V)

Repérage des patrons sémantico-syntaxiques récurrents

✓Extraction de configurations intégrant des contraintes sur les relations syntaxiques et les traits sémantiques et comportant un MD

✓Informations sur la combinatoire des MD et de leur rôle structurant à travers des routines

MÉTHODOLOGIE (VI)



À partir des réalisations lexicales en corpus →

Généralisation d'un patron syntactico-sémantique :

(SUBJ) Verbe$_{\text{description}}$ Adverbe$_{\text{énumération}}$ (OBJ Nom$_{\text{réalisation}}$)

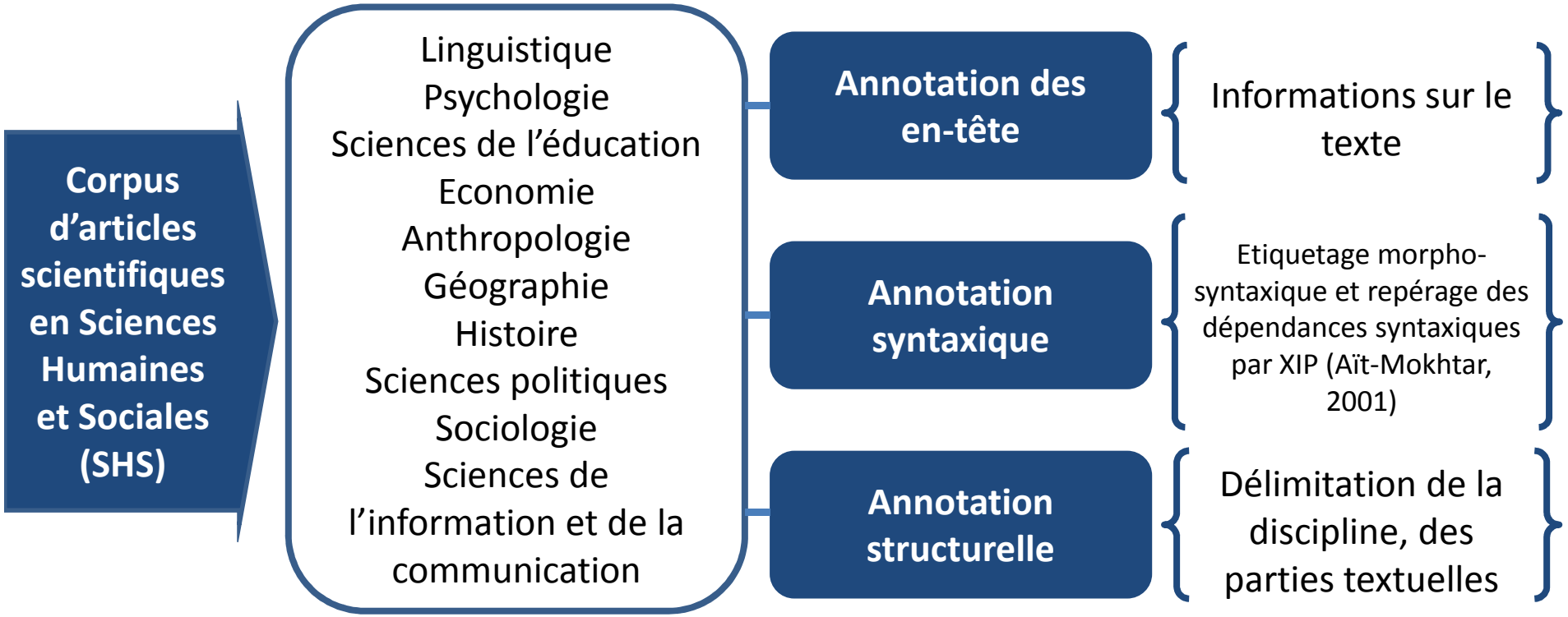
Verbe$_{\text{description}}$

décrire
détailler
présenter
...

Nom$_{\text{réalisation}}$

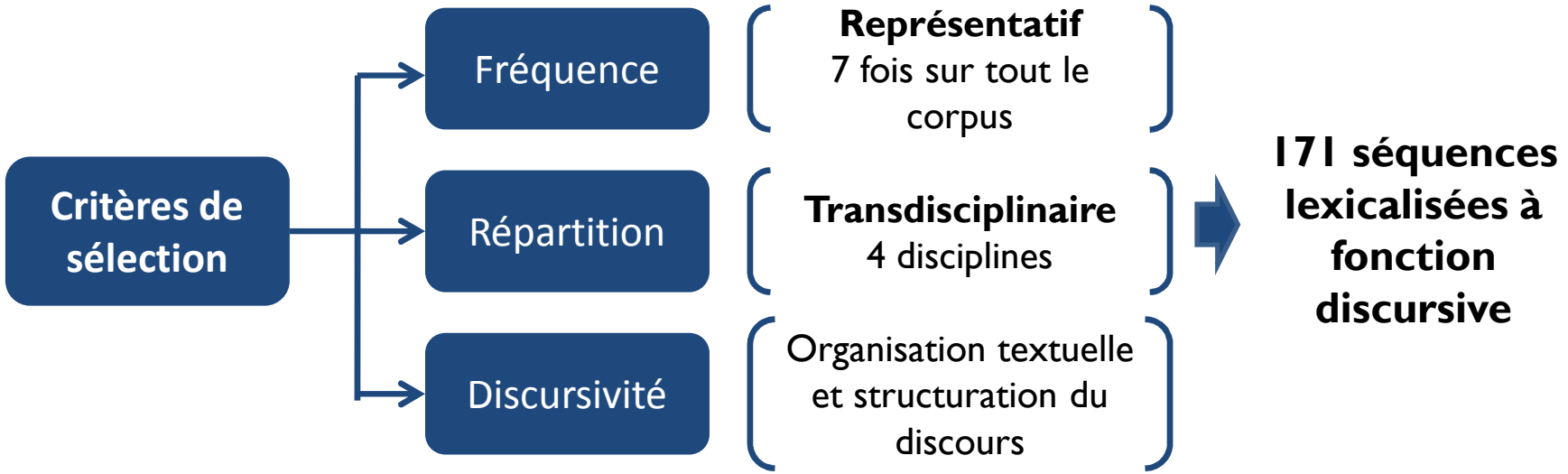
recueil
constitution
élaboration
...

MÉTHODOLOGIE (VII)



499 articles ≈ 4 920 000 mots

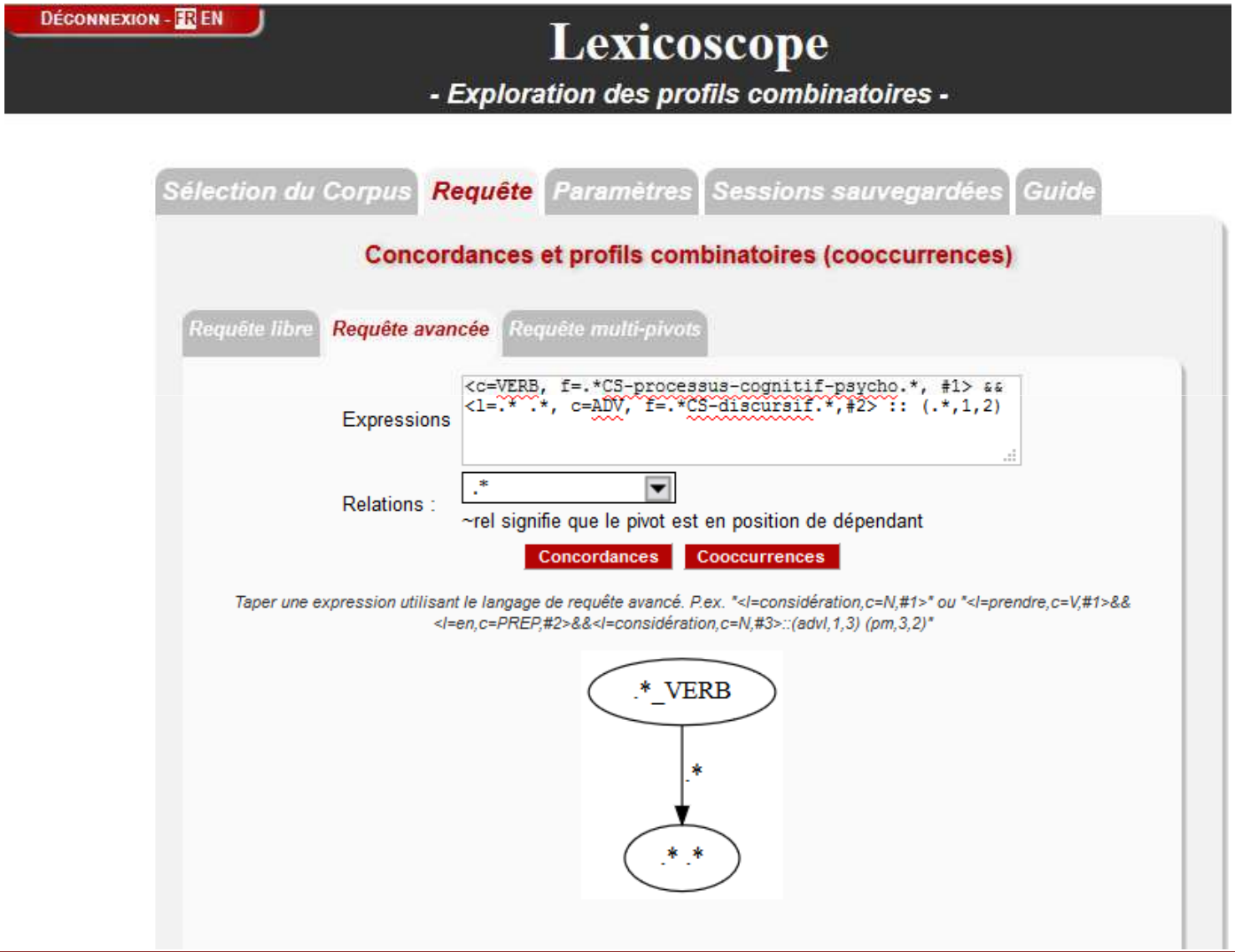
MÉTHODOLOGIE (VIII)



TYPOLOGIE DES SÉQUENCES LEXICALISÉES À FONCTION DISCURSIVE

Fonction	Types de marqueurs	Exemples	
Métadiscursive	Enchaînement thématique ✓Addition ✓Topicalisation ✓Structuration ✓Enumération ✓Conclusion	 Par ailleurs, de surcroît A propos de, en ce qui concerne D'une part, d'autre part etc. Tout d'abord, en premier lieu En conclusion, pour conclure	**Organisation textuelle**
	Reformulation	C'est-à-dire, en somme	
	Exemplification	Par exemple, en particulier	
	Comparaison	D'autant plus, à l'instar de	
Argumentative	Causalité	En effet, en conséquence	**Raisonnement logique**
	Concession	D'ailleurs, en revanche	
	Opposition	Au contraire, par opposition	
	Condition	A condition que, à supposer que	
	Finalité	En vue de, à cet effet	

VERBES + MD (I)


DÉCONNEXION - FR EN
Lexicoscope
- Exploration des profils combinatoires -
Sélection du Corpus
Requête
Paramètres
Sessions sauvegardées
Guide
Concordances et profils combinatoires (cooccurrences)
Requête libre
Requête avancée
Requête multi-pivots
Expressions
<c=VERB, f=.*CS-processus-cognitif-psycho.*, #1> &&
<l=.* .*, c=ADV, f=.*CS-discursif.*,#2> :: (.*,1,2)
Relations :
.*
~rel signifie que le pivot est en position de dépendant
Concordances
Cooccurrences
Taper une expression utilisant le langage de requête avancé. P.ex. *<l=considération,c=N,#1>* ou *<l=prendre,c=V,#1>&&
<l=en,c=PREP,#2>&&<l=considération,c=N,#3>::(advl,1,3) (pm,3,2)*
.*_VERB
.*
.* .*

VERBES + MD (II)

Nombre total d'occurrences 113
Dispersion 10

Requêtes

- Requête : <c=VERB, f=.*CS-processus-cognitif-psycho.*, #1> && <l=.* .*, c=ADV, f=.*CS-discursif.*,#2> :: (.*,1,2)

Show 10 entries Search:

Identifiant	Contexte gauche	Pivot	Contexte droit
XML_ling.xml-s103	L' intérêt de ces phrases ambiguës (4) , couramment utilisées par les grammairiens est , selon nous , essentiellement didactique , elles permettent de	montrer	d'une part , que le déplacement du complément , et sa position en tête de phrase , induisent un changement de sens et une insertion dans un contexte textuel ou scénique (Riegel , Le Goffic) , d'autre part que le G Prép .
XML_histoire.xml-s117	Ce privilège moderne accordé à l' orientation sexuelle n' est bien évidemment pas absolu , soudain et total , et met plusieurs décennies à s' imposer , notamment dans les classes populaires , comme le	montrent	très bien les écrits de Daniel Guérin par exemple [26] .
XML_scinfo.xml-s117	Cela	permet	, par exemple , de conserver des sessions (après une première authentification , l' utilisateur reste connecté à un site privé) , de gérer des « paniers électroniques » , etc .
XML_sciencespolitiques.xml-s12	Ainsi a -t-on particulièrement prêté attention à tout ce qui	souligne	dans ces théories les supposées causes « organiques » des différences entre humains , en particulier les arguments et formules qui visent à localiser et à enfermer dans la « nature » d' un corps ou d' une lignée - classes , races , criminels , prostituées , génies , aliénés , etc .

COOCCURRENTS

Cooccurrents	Fréquence	Répartition
Montrer par exemple	25	8
Montrer par ailleurs	22	10
Montrer en particulier	17	8
Présenter dans un premier temps	10	6
Souligner d'ailleurs	8	7
Montrer tout d'abord	8	4

APPRENTISSAGE SUR CORPUS

- Introduction du corpus pour l'enseignement/apprentissage des MD :
  - Absence d'utilisation du corpus en classe de langue (Gilquin & Granger, 2010).
  - Intérêt de l'ASC Apprentissage Sur Corpus (Boulton, 2011) : accéder aux données authentiques, favoriser une meilleure prise de conscience linguistique (Cavalla & Loiseau, 2013).

- Intérêt d'une annotation sémantique :
  - Accès onomasiologique (accès par le sens aux éléments lexicaux) : propice pour développer des compétences rédactionnelles
  - Fréquence des MD
  - Mieux voir la structure du texte grâce aux MD

EXEMPLE

MD d'introduction thématique
MD + groupe nominal
***Fréquence*: ***

MD d'énumération
Fréquence:* *

Ces trois expérimentations nous ont permis d'évaluer le modèle de représentation du texte, le langage SEXTANT et les interfaces de navigation qui sont proposées aux utilisateurs. En ce qui concerne le modèle de représentation du texte, les résultats confirment nos hypothèses : les différentes relations convoquées, syntagmatiques, syntaxiques, sémantiques et discursives ont pu être facilement représentées tout en conservant une bonne lisibilité. Il convient néanmoins de pondérer ce constat en remarquant que ces trois expérimentations ne convoquent pas de représentations intratextuelles complexes. En ce qui concerne le langage SEXTANT, les résultats sont là aussi satisfaisants, puisque toutes les opérations de navigation demandées par les experts ont pu être exprimées. Néanmoins, certaines limitations du langage de conditions sur lequel s'appuie le langage SEXTANT sont apparues. Notamment, il n'est pas possible d'exprimer une condition sur la position relative d'une UT cible dans le texte, autrement qu'en faisant référence à l'UT source. Si aucune des expérimentations n'a nécessité l'expression de telles conditions, il n'en reste pas moins que ceci pourrait constituer une limitation pour certains développements futurs. Quant à la gestion des interactions, le menu actuel de choix des opérations de navigation satisfait les besoins essentiels, mais sa position dans le texte ne peut pas être paramétrée, il se positionne toujours sur l'UT source, ce qui rend la lecture quelquefois difficile.

Ces expérimentations nécessitent l'expression d'opérations de navigation qui convoquent des expertises très différentes. L'expérimentation NaviLire a bénéficié de la longue expérience pédagogique et des réflexions théoriques de L. Lundquist sur l'enseignement de la linguistique textuelle. Nous disposions ainsi d'une documentation abondante (exemples d'exercices sur support imprimé, ouvrages publiés) que nous avons aisément transcrite sous la forme d'opérations de navigation. De même, les structures textuelles ont été balisées (manuellement) par L. Lundquist à partir de ces mêmes exercices. Pour l'expérimentation sur le résumé, nous avons déjà mentionné les travaux sur lesquels nous nous sommes appuyés et il faut simplement remarquer que les structures textuelles sont repérées automatiquement par un système dédié au filtrage d'informations. En revanche, les opérations de navigation, même si elles sont issues de divers travaux ont été spécifiées empiriquement et une évaluation en usage reste à faire. Enfin, l'expérimentation sur la navigation dans le roman « Madame Bovary » est la plus prospective mais elle a le mérite de montrer que des relations de types syntaxiques et ontologiques peuvent être représentées dans le formalisme proposé par NaviTexte.

Javier Couto, Jean-Luc Minel, Navigation textuelle : représentation des textes et des connaissances. Traitement Automatique des Langues, 47-2, 2006.

1
2
3
Evaluation de trois expérimentations

1'
2'
3'
Fondements théoriques de trois expérimentations

CONCLUSION & PERSPECTIVE

- Utilité des corpus pour :
  - Des analyses linguistiques :
    - ✓ Mettre en évidence les propriétés linguistiques
    - ✓ Repérer le fonctionnement des unités lexicales
  - Des activités didactiques : exemples authentiques

- Une annotation sémantique semi-automatique:
  - Rôle de l'enseignant : sélection des données, organisation de l'apprentissage.

- Gestion du problème des polylexicaux pour l'analyseur
- Mise en place d'un outil pour une visualisation explicite

RÉFÉRENCES BIBLIOGRAPHIQUES

- Adam, J.-M. (1989). Aspects de la structuration du texte descriptif : les marqueurs d'énumération et de reformulation. *Langue française*, (81), 59-98.
- Aït-Mokhtar, Salah, Jean-Pierre Chanod, and Claude Roux. "Robustness beyond shallowness: incremental deep parsing." Natural Language Engineering 8.3 (2002): 121-144
- Boulton, A. & Tyne, H. (2014): *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues*. Paris: Didier.
- Cavalla, C. (à paraître). Collocations transdisciplinaires dans les écrits de doctorants FLS/FLE. In F. Sitri & F. Rinck (Éd.), *Former à l'écrit universitaire*. Paris : Linx.
- Cavalla, C., & Loiseau, M. (2013). Scientext comme corpus pour l'enseignement. In A. Tutin & F. Grossmann (Éd.), *L'écrit scientifique : du lexique au discours. Autour de scientext*. Rennes : Presses Universitaires de Rennes, 163-180.
- Drouin, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. Revue française de linguistique appliquée, 12(2),
- 45-64
- Hatier, S., Tutin, A., Jacques, M-P., Jacquey, E. & Kirster, L. (2014), Catégorisation sémantique des noms simples du lexique scientifique transdisciplinaire. ACFAS, Colloque *étude de lexiques à vocation particulière.*
- Legallois, D., & Tutin, A. (2013). Présentation : Vers une extension du domaine de la phraséologie. *Langages*, *1*(189), 3-25.
- Paillard, D. & Vu, T.-N. (2012): *Inventaire raisonné des marqueurs discursifs du français. Description. Comparaison. Didactique.* Paris: AUF.
- Phal, A. (1971): *Vocabulaire général d'orientation scientifique*. Paris: Crédif.
- Steuckardt, Agnès/Niklas-Salminen, Aïno (2005): "Les marqueurs de glose." *Langues et langage* 12/1. Aix-en-Provence: Publications de l'Université de Provence.
- Tran, T.-T.-H. (2014). *Développement d'une aide à l'écrit scientifique. Description de la phraséologie scientifique et réflexion didactique pour l'enseignement à des étudiants non natifs.* (Thèse de doctorat). Grenoble: Université Grenoble Alpes.
- Tutin, A., & Falaise, A. (2013). Multiword expressions in scientific discourse : a corpus-driven database. In *Actes de eLex*, Tallinn, Estonie, Talinn, Estonie, 2013.
- Tutin, A. (2007) : "Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques." *Actes de Traitement Automatique des Langues Naturelles (TALN)*: 283-292.

Merci