# Lexique transdisciplinaire et structure des articles scientifiques

Gwendoline Bloquet, Agnès Tutin

LIDILEM
Université Grenoble 3

Colloque « Sciences et Ecritures »
13-14 mai, LASELDI, Besaçon

# CADRE

- **Le but de l'étude :**
  - Proposer une **étude linguistique** du lexique transdisciplinaire des articles scientifiques dans une perspective d'aide à la rédaction.
    - Perspective de Traitement Automatique du langage dans le cadre d'un projet piloté par le LIDILEM (« Acquisition automatique de traductions d'expressions semi-figées pour l'élaboration d'outils d'aide à la rédaction scientifique et technique », Projet EMERGENCE Rhône-Alpes).
    - Les expressions semi-figées étudiées ➔ collocations du type *faire une hypothèse* ou *préconiser une méthode*.
  - Observer la **répartition du lexique nominal selon l'intention communicative** à partir d'un balisage XML.

# CADRE

- **Hypothèse :**
  - **Forte corrélation entre la nature du lexique nominal employé et l'intention communicative**
    - Cette corrélation, si elle est établie, permettra d'orienter la conception des outils d'aide à la rédaction.
- **Méthode** :
  - Constitution d'un **corpus de 21 écrits** et découpage logique de la structure (balisage XML).
  - **Etude de corpus** (21 écrits) linguistique fondée sur des critères syntaxiques et sémantiques. .

# 1. Constitution et traitement du corpus

- **Constitution du corpus**
  - **Le corpus est constitué de 21 textes:**
    - **15 articles**
    - **4 rapports**
    - **2 thèses/chapitres de thèse**
  - **Taille 194114 mots au total**
  - **Domaines:**
    - **Traitement Automatique du langage (9)**
    - **Traitement de l'information (1)**
    - **Traitement de l'image (1)**
    - **Médecine (6)**
    - **Génétique (1)**
    - **Linguistique (2)**
    - **Épidémiologie (1)**
  - **Contrainte :** Textes français ↔ anglais.

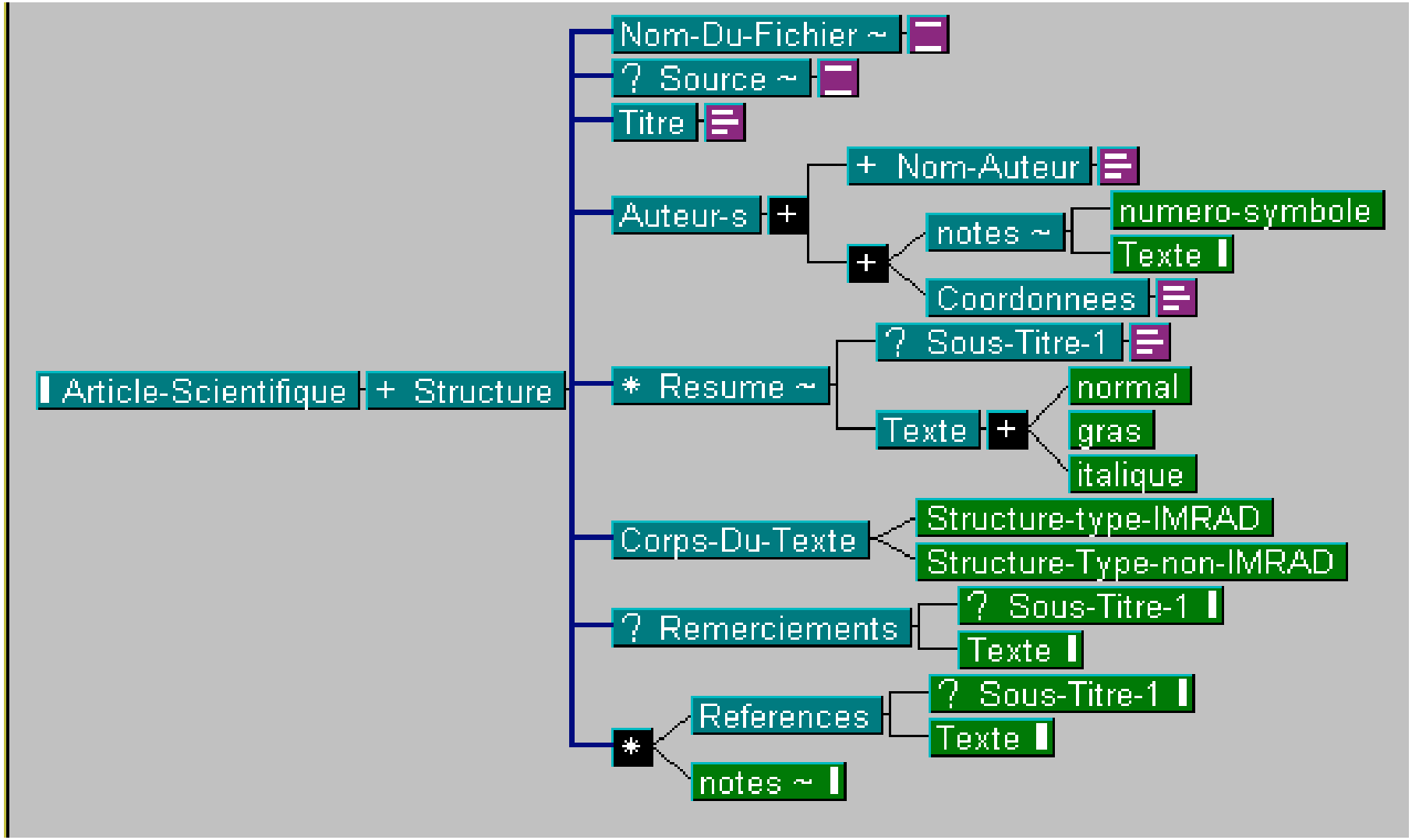
# Balisage logique du document

- **Un balisage inspiré du projet KIAP** (Kjersti Fløttum, Université de Bergen, Institut d'Études romanes, Norvège)
  - Projet contrastif, qui porte sur l'identité culturelle (existe-t-elle?) dans le discours académique (plus précisément dans les articles de recherche).
  - Le cadre théorique se définit par des études sur les modalités, le métadiscours, les citations/le discours rapporté, la polyphonie linguistique, la sémantique lexicale et la sémantique interprétative textuelle.
- **Balisage XML utilisé dans notre projet**
  - Intérêt d'un tel balisage : Observer la répartition du lexique selon la structure logique du document.

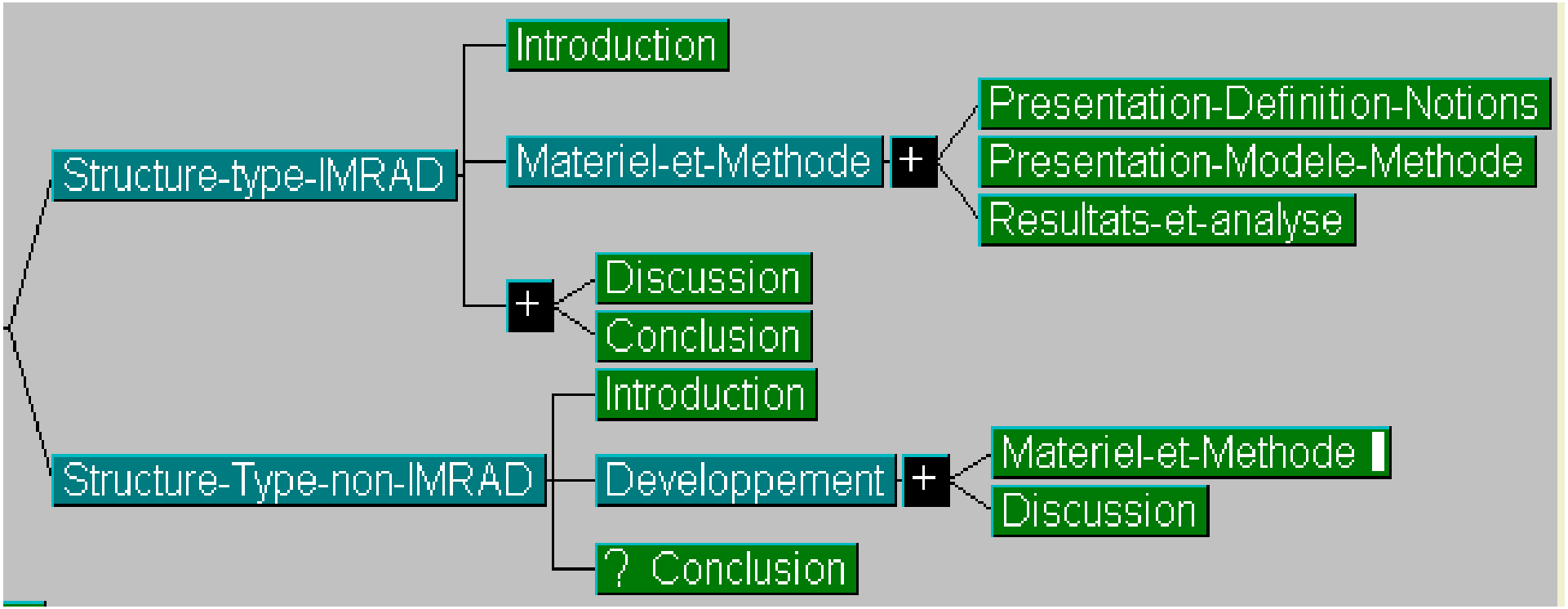
- **Balisage logique du document**
  - **La DTD (Description/Définition de Type de Document)**
    - *Définition*: Elle permet d'expliquer au système comment seront structurés les documents qui relèvent de la DTD en question.

    ➔ C'est en quelque sorte la syntaxe du document.
    - Création d'une DTD sous forme arborescente
    - Présentation sommaire de la DTD
      - Deux structures de l'article scientifique :
        - IMRAD ( ou IMRED en français) : Plan international en quatre parties : Introduction, Matériel et Méthodes, Résultats et Discussion
        - Non-IMRAD: structure dans laquelle on peut avoir les éléments de la structure IMRAD dans n'importe quel ordre

# Présentation globale de la DTD



# Présentation détaillée des structures IMRAD et Non-IMRAD



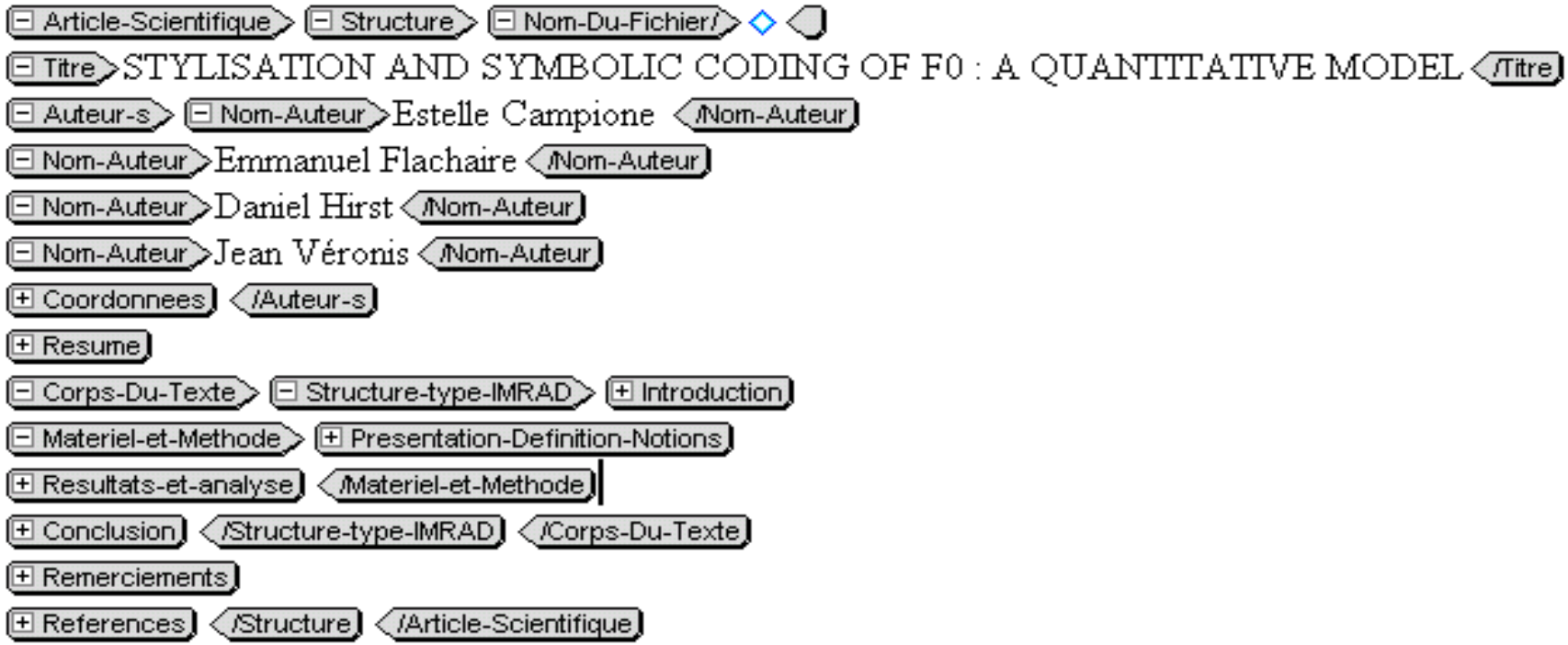

1. Constitution et traitement du corpus

# Balisage logique du document

## Balisage des textes sous un éditeur de XML

## ➔ XMetal

### Exemple d'article balisé:



# 2. Etude linguistique du corpus

- **Porte sur le lexique transdisciplinaire des écrits scientifiques**
  - **Lexique commun aux écrits scientifiques** :

➔ Ne relève pas de la langue générale, ni des langues de spécialité.

  - Apparaît **plus large que le lexique méthodologique** (ne concerne pas que l'exposition de la méthode).
  - Est **moins large que le lexique épistémique** (qui englobe aussi le lexique de spécialité).

- **Travaux en anglais dans le domaine de l'« English for academic purposes » (Coxhead 1998, 2000)**
  - **Etude de linguistique de corpus sur un corpus d'écrits « académiques** » :
    - 3,5 millions de mots
    - Domaines variés dans une perspective pédagogique.
  - **Constitution de l'« Academic Word List »**

    Critères statistiques ➔ Sélection des mots qui :
    - apparaissent plus de 100 fois
    - ne sont pas les plus fréquents de l'anglais
    - sont présents dans 10 sous-domaines :
      - ➔ **3100 mots appartenant à 570 familles morphologiques**.
  - Inconvénient des critères statistiques ➔ produisent des listes qui ne rendent **pas compte de la polysémie et des cooccurrences lexicales.**

➢ **Approche intéressante mais qui doit être complétée par une étude sémantique et une prise en compte du contexte.**

- **Sélection des noms transdisciplinaires**
  - **A partir d'un calcul de fréquences** des noms les plus fréquents (une quarantaine de noms) (sur un corpus de 190 000 mots).
  - **Sélection manuelle** à l'aide des contextes (à l'aide des concordances).

associée à nos 145 paires de FA. Cette méthode a donné de bien meilleurs résultats, comme on p
modèle de langue markovien caché. Cette méthode a fait l'objet d'une implantation réelle : le p
éralisation : La généralisation de la méthode à plus de deux sources est évidente et se justi
r les hypothèses considérées de OSi La méthode a pour inconvénient de dégrader la qualité de d
ndant. Nous présentons ci-dessous cette méthode appliquée dans le cas des trois sources suivant
igham Young et Xerox PARC6. La première méthode automatique d'alignement de textes parallèles a
ions concernant la comparaison de cette méthode avec celle de (129) 92 Chapitre III : Soluti
lités et pour lequel nous proposons une méthode capable d'en déduire un jeu de masses, correspo
7 Conclusion : Nous avons proposé une méthode capable de combiner, de façon globale, plusieur
chain souschapitre, nous présentons une méthode capable de déduire un jeu de masses dont le cal
nt ainsi déduit, nous proposons ici une méthode capable de déterminer un jeu de masses lui corr
ales obtention, nous proposons ici une méthode capable de trouver un jeu de masses corresponda

  - **Vérification à partir des concordances** :

    Exemple : exclusion de *technologie,* secteur d'activité plus qu'ensemble de procédures.

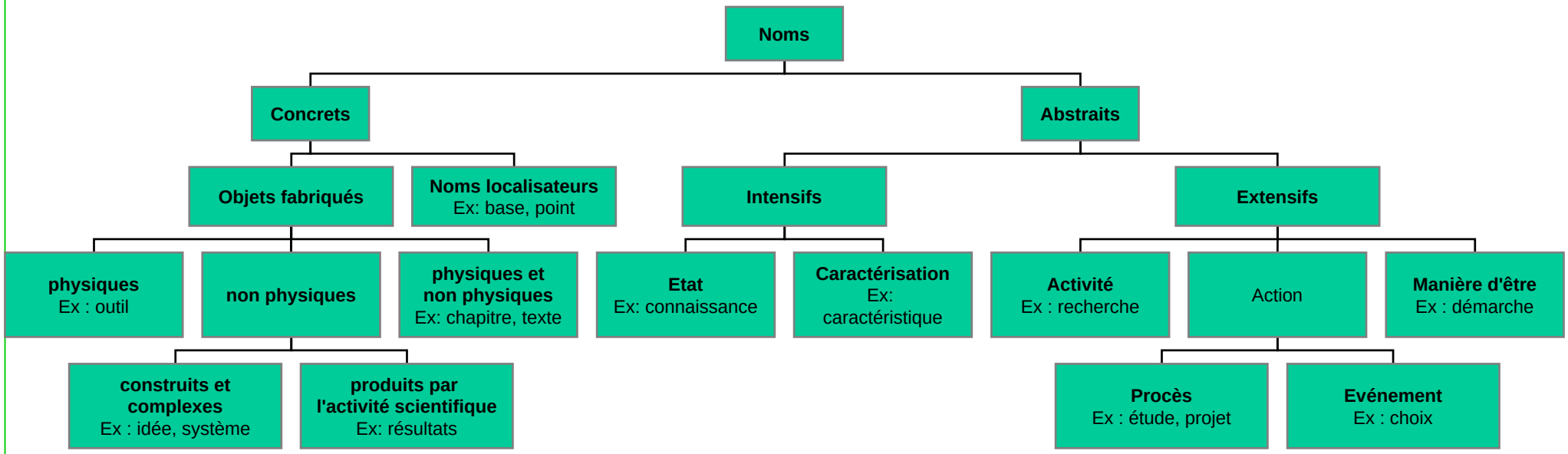
# Etude sémantique du lexique transdisciplinaire

- Basée sur des **critères linguistiques**
  - Critères formels et reproductibles
    - Problèmes des taxonomies de type ontologique comme *Wordnet*

    (➔ Dictionnaire informatisé dont l'unité de base est le concept et non le mot.)
  - Doit permettre de prédire les associations lexicales
    - Exemple: Collocations du type *faire, formuler une hypothèse.*

- S'inspire de la **classification de Flaux et Van de Velde (2002) Ophrys)**
  - Classification linguistique, largement basée sur les propriétés syntaxiques, qui présente des critères explicites.
    - Exemple : les procès sont des noms d'action qui se combinent avec le support *faire* et des variantes aspectuelles marquant une étape de l'action (*faire une étude, démarrer/entamer une étude, clore/achever une étude*).
  - Classification large (mais comportant des lacunes et privilégiant les noms mettant en jeu des agents humains)

# Classification des noms transdisciplinaires




2. Etude linguistique du corpus

# 3. Répartition du lexique transdisciplinaire

- **Répartition du lexique selon la structure fonctionnelle des éléments de l'article.**
- **Etude sur un sous-ensemble de noms significatifs :**
  - **noms de procès : *analyse, application, développement, étude, recherche, traitement, utilisation.***
  - **noms d'objets construits « élaborés » : *analyse, application, idée, hypothèse, méthode, modèle, outil, système, technique, technologie, théorie, traitement.***
  - **noms d'objets produits (par l'activité scientifique) : *cas, donnée, exemple, problème, question, résultat*.**
- **Etude sur un sous-ensemble du corpus : 11 documents (61 500 mots) dont la structure logique a été découpée.**

# Méthode

- **Repérage de fréquences des différents types de noms selon la structure logique (désambiguïsation manuelle de certains noms).**
- **Répartition selon les structures suivantes :**
  - Introduction
  - Présentation- Modèle - Méthode
  - (Résultats et) analyse
  - Discussion
  - Conclusion
- **Découpage fonctionnel peu fin, de nombreux textes ne présentant pas une structure IMRAD.**

# Résultats

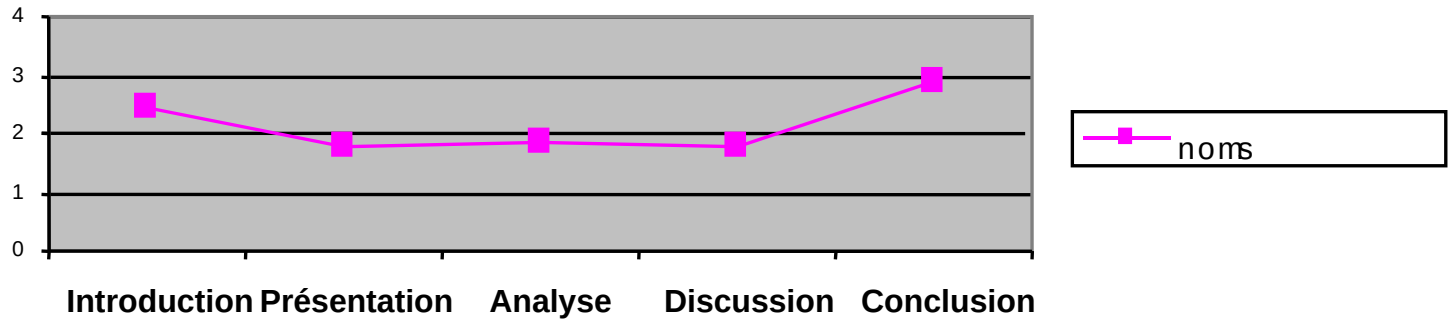
**Répartition des noms selon la structure logique et selon leur type**

| | **Introduction** | **Présentation-modèle – méthode** | **(Résultats et) Analyse-** | **Discussion** | **Conclusion** |
|---|---|---|---|---|---|
| **N_procès** | 32 (% 0,4) | 110 (% 0,3) | 69 (% 0,54) | 8 (% 0,3) | 30 (% 0,72) |
| **N_élaborés** | 72 (% 1,1) | 333 (% 0,94) | 78 (% 0,61) | 13 (% 0,5) | 58 (% 1,4) |
| **N_produits** | 58 (% 0,9) | 205 (% 0,57) | 90 (% 0,71) | 28 (% 1,0) | 32 (% 0,77) |
| **Total** | 162 (% 2,45) | 648 (% 1,8) | 237 (% 1,86) | 49 (% 1,8) | 120 (% 2,9) |
| **Nbre de mots** | 6593 | 35 562 | 12 675 | 2 602 | 4132 |

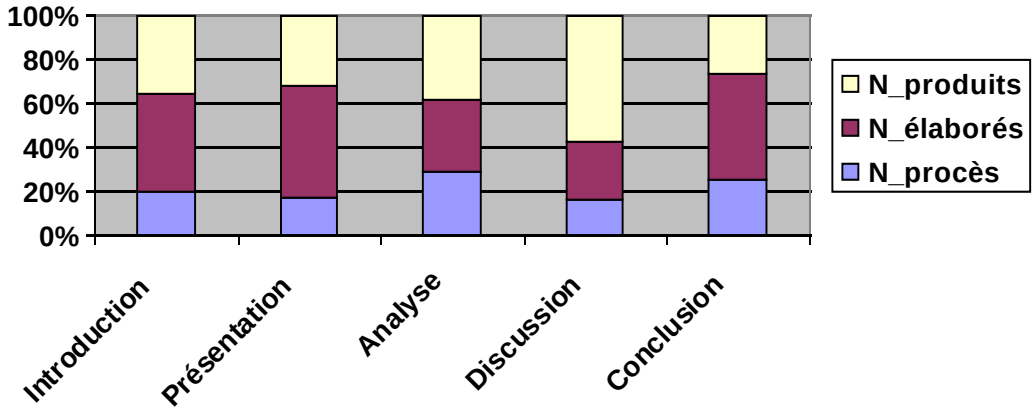
3. Répartition du lexique transdisciplinaire

# Résultats

**Répartition des noms selon la structure logique**




**Répartition des types de noms selon la structure logique**




3. Répartition du lexique transdisciplinaire

# Analyse

- **Densité en noms « transdisciplinaires » plus importante dans l'introduction et la conclusion (accent sur la démarche, sur la méthodologie?).**
- **Les différents types de noms apparaissent (en proportion non négligeable) dans les différentes structures du document.**
- **Quelques tendances se dégagent :**
  - **Les noms de procès (action) sont plus nombreux dans les parties d'analyse et de conclusion.**
  - **Les noms de concepts élaborés (*méthode, hypothèse, …)* apparaissent surtout dans l'introduction, la présentation de la méthode et la conclusion.**
  - **Les noms correspondant à des « produits » de l'activité scientifique (*résultats, cas, données* …) sont surreprésentés dans les parties Analyse et Discussion.**

# Analyse

- **Néanmoins, la corrélation entre le type de nom et la structure logique n'est pas flagrante.**
- **Explications :**
  - **Un corpus trop hétérogène, contenant trop de textes de structures différentes?**
    - **L'analyse de corpus de structures IMRAD produirait peut-être des corrélations plus criantes.**
  - **Un découpage structurel trop grossier et trop rigide, ne rendant pas compte des intentions communicatives du scripteur?**
    - **Le découpage logique est basé sur les parties définies par l'auteur, peu fines et ne présentant pas toujours un but communicatif unique.**
  - **Les intentions communicatives ne s'expriment-elles pas plutôt à travers les verbes?**
    - **Les notions nominales, plus stables, se déclinent dans plusieurs intentions communicatives.**

# Conclusion

- **Conséquences pour un système d'aide à la rédaction pour différents types d'articles, non nécessairement de structure IMRAD :**
  - **Le choix lexical en fonction de l'intention communicative ne sera pas en priorité guidé par les noms, mais peut-être plutôt par les verbes.**
- **Le balisage logique d'un corpus hétérogène d'articles est difficile et rend difficilement compte de l'intention communicative.**
- **Nécessité d'études sur le lexique verbal et sur des corpus plus conséquents.**

# Bibliographie


- COXHEAD, Averil (2002) 'The Academic Word List: A Corpus-based Word List for Academic Purposes' in *Teaching and Language Corpora (TALC) 2000 Conference Proceedings* Rodopi: Atlanta.
- COXHEAD, Averil (2000) 'A new academic word list' *TESOL Quarterly*,34(2): 213-238.
- FLAUX N., VAN DE VELDE D. (2002), *Les noms en français*, Paris, Ophrys.
- HESLOT, J., (1978), « « we found that... » Approche du discours scientifique en anglais », *Langue et Société*, supplément au n° 5, Paris, Maison des Sciences de l'Homme, pp. 16-19
- HESLOT, J., (1980), « La formation des chercheurs à l'expression scientifique écrite », in *Langue et Société*, supplément au n° 12, Paris, Maison des Sciences de l'Homme, pp. 35-40
- HESLOT, J., (1983), « Récit et commentaire dans un article scientifique », *D.R.L.A.V.*, n° 29, Paris, Centre de recherche de l'université de Paris VIII
- NUCHEZE, V. de, (1991), « Les typologies à la lumière d'un genre hybride : le discours de recherche », in BRONCKART, J.P. et al, (dir.), *Etudes de linguistique appliquée*, n° 83, Paris, Didier érudition, pp.101-115
- NUCHEZE, V. de, (1998), « Approche pragmatico-énonciative du discours de recherche (à l'usage des apprentis-chercheurs), in DABENE, M., REUTER, Y., (coord.), *Pratiques de l'écrit et modes d'accès au savoir dans l'enseignement supérieur, Lidil,* n° 17, U. Stendhal, Grenoble