Le genre comme point d'accès au document

Analyse comparée de textes scientifiques en mécanique et linguistique

Viviane Clavier

Laboratoire Gresec

Université Stendhal (Grenoble)

Intérêt du genre pour la recherche d'information

- Texte intégral, utilisateur
- Textes, normes, genres, interprétation
- Terminologie *vs* genre textuel et accès à l'information
- Domaine, genre et classification de textes

Plan

- Approche descriptive
  - Genre et discours ; genre et domaine
  - Caractérisation morphosyntaxique
- Application pour la recherche d'information
  - classification
  - indexation

Le genre à l'épreuve des discours et des domaines

Contexte de l'étude : collaboration

Poudat Céline

Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres.

Thèse de doctorat soutenue à l'Université d'Orléans, juin 2006.

Définition du genre

« le genre est un niveau normatif de contraintes et de prescriptions positives ou négatives régulant la production et l'interprétation d'un texte. Puisque tout texte relève d'un genre, et que tout genre relève d'un discours, le genre permet de relier les textes aux discours »

Méthodologie retenue

- Description morphosyntaxique des textes (Malrieu et Rastier, 2001)
- Méthode de profilage (Habert, 2000)

Outils retenus

- Etiqueteurs
  - Cordial® : comparaison article scientifique / autres discours
  - TnT Tagger entraîné sur un corpus de linguistique avec un jeu d'étiquettes dédié au genre de l'article linguistique.
- Statistiques descriptives et multidimensionnelles (ACP)

Corpus de mécanique

- 49 textes scientifiques qui sont des actes de conférence extraits du XV^{e} Congrès Français de Mécanique (2001)
  - *Aérodynamique*
  - *Biomécanique*
  - *Hydrodynamique*

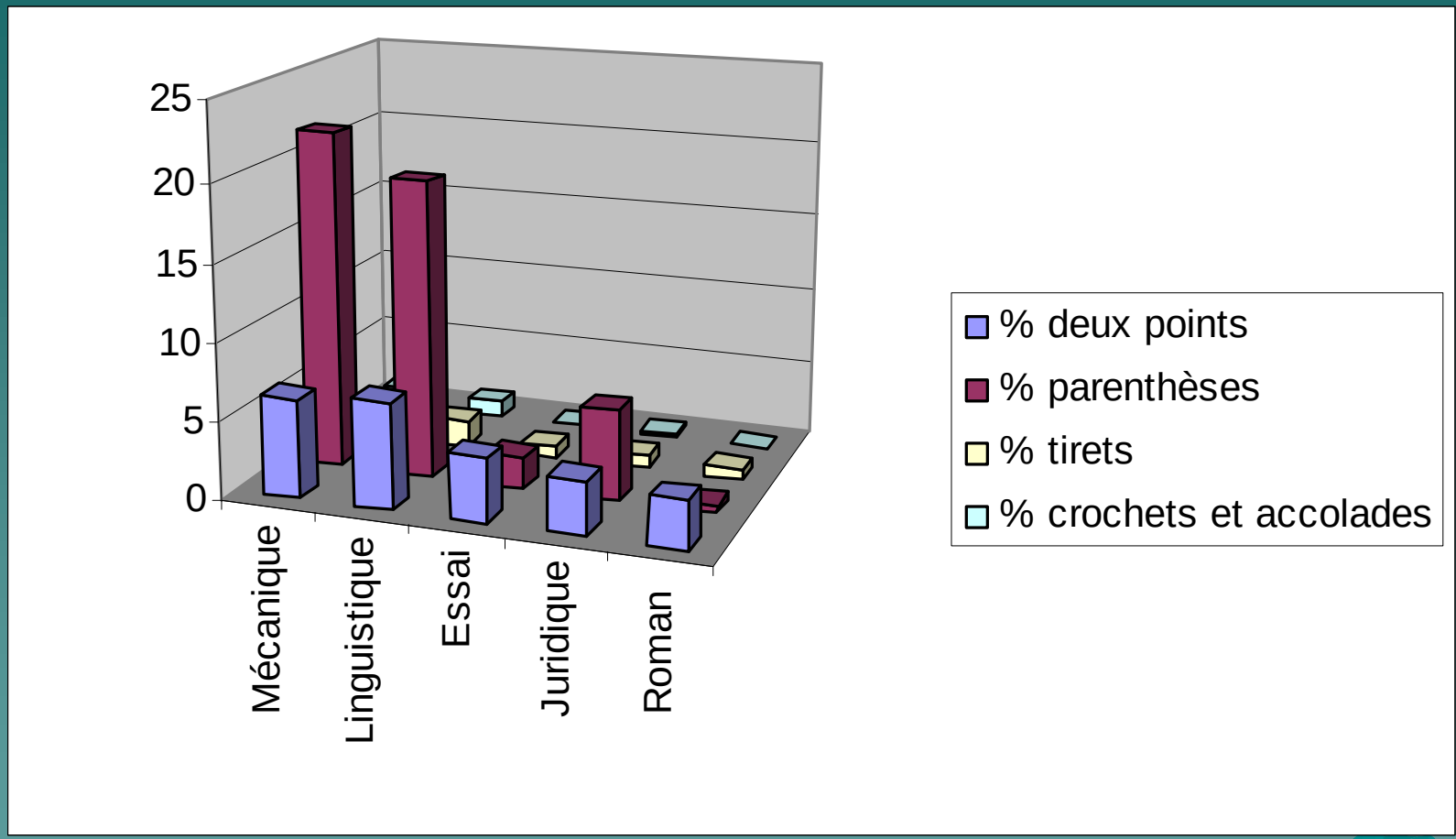
Corpus de linguistique (Poudat, 2006)

- 224 textes : articles scientifiques français de revues linguistiques
  - Les *Cahiers de Praxématique*
  - Les *Cahiers du CIEL*
  - *Langue française*, etc.

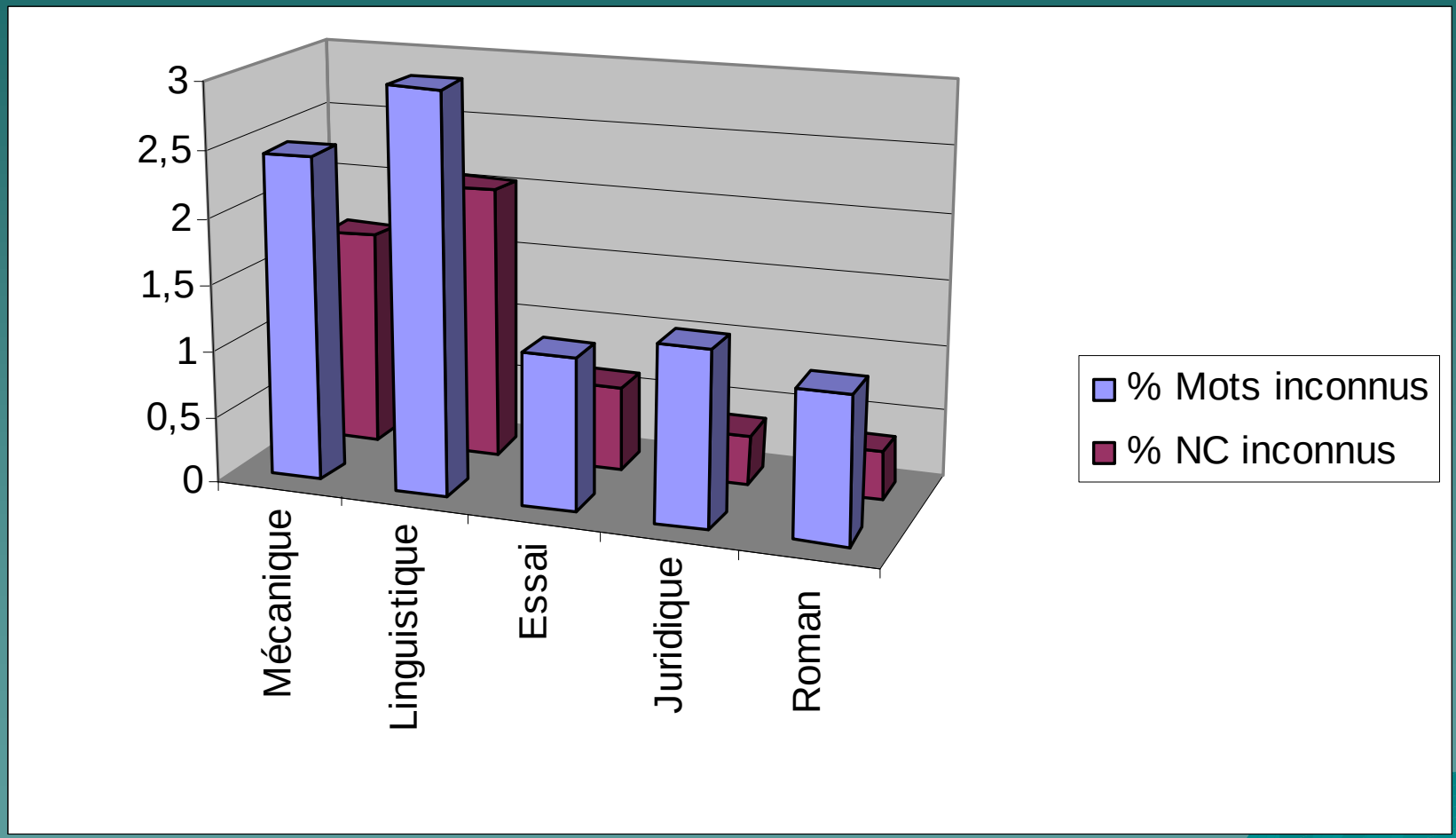
Comparaison du discours scientifique *versus* autres discours

- Confrontation du corpus d'articles scientifiques à un corpus de référence (romans, textes juridiques, essais)
  - source D. Malrieu
- Cordial® (Synapse Développement)

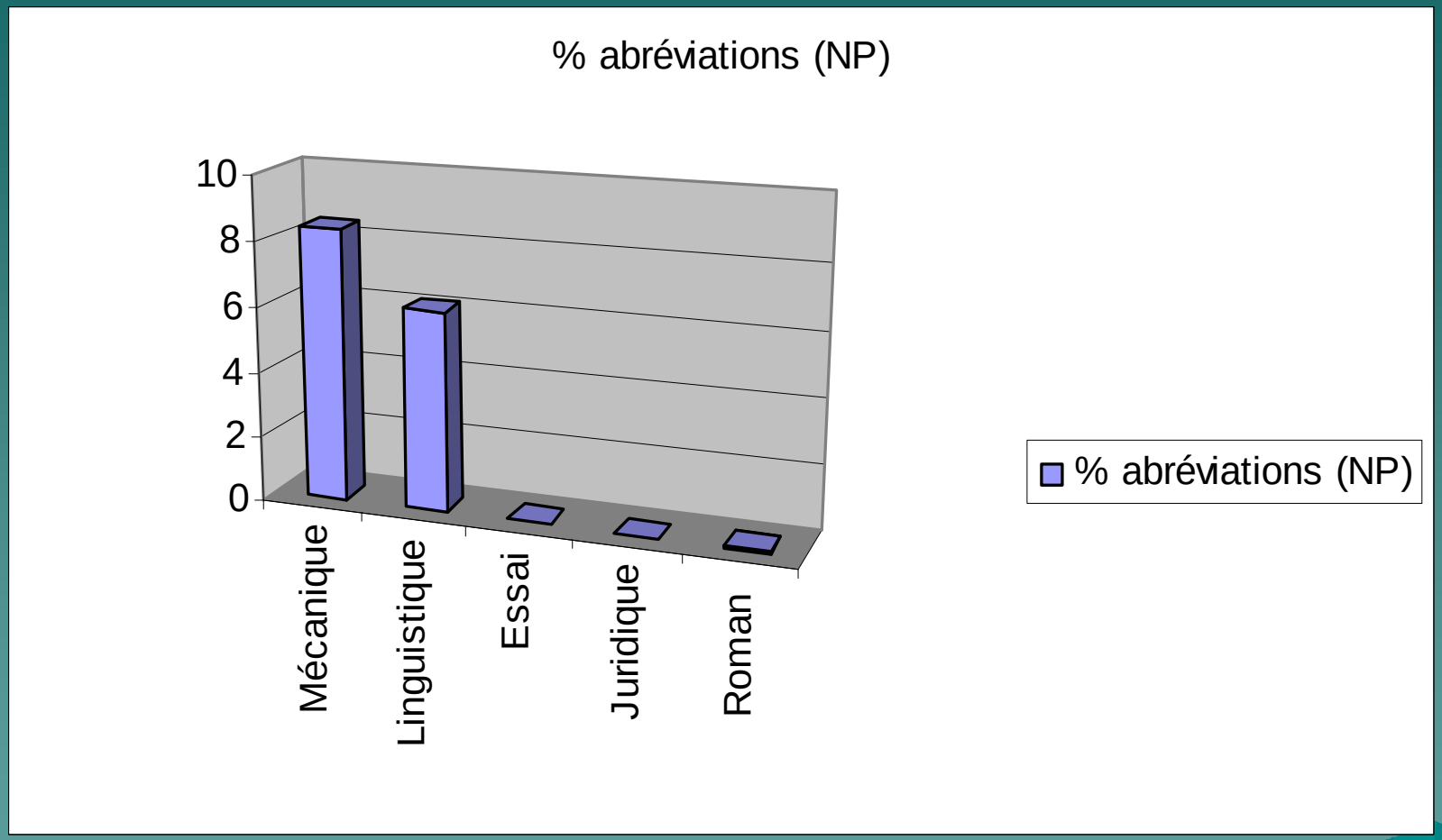
Résultats : ponctuation



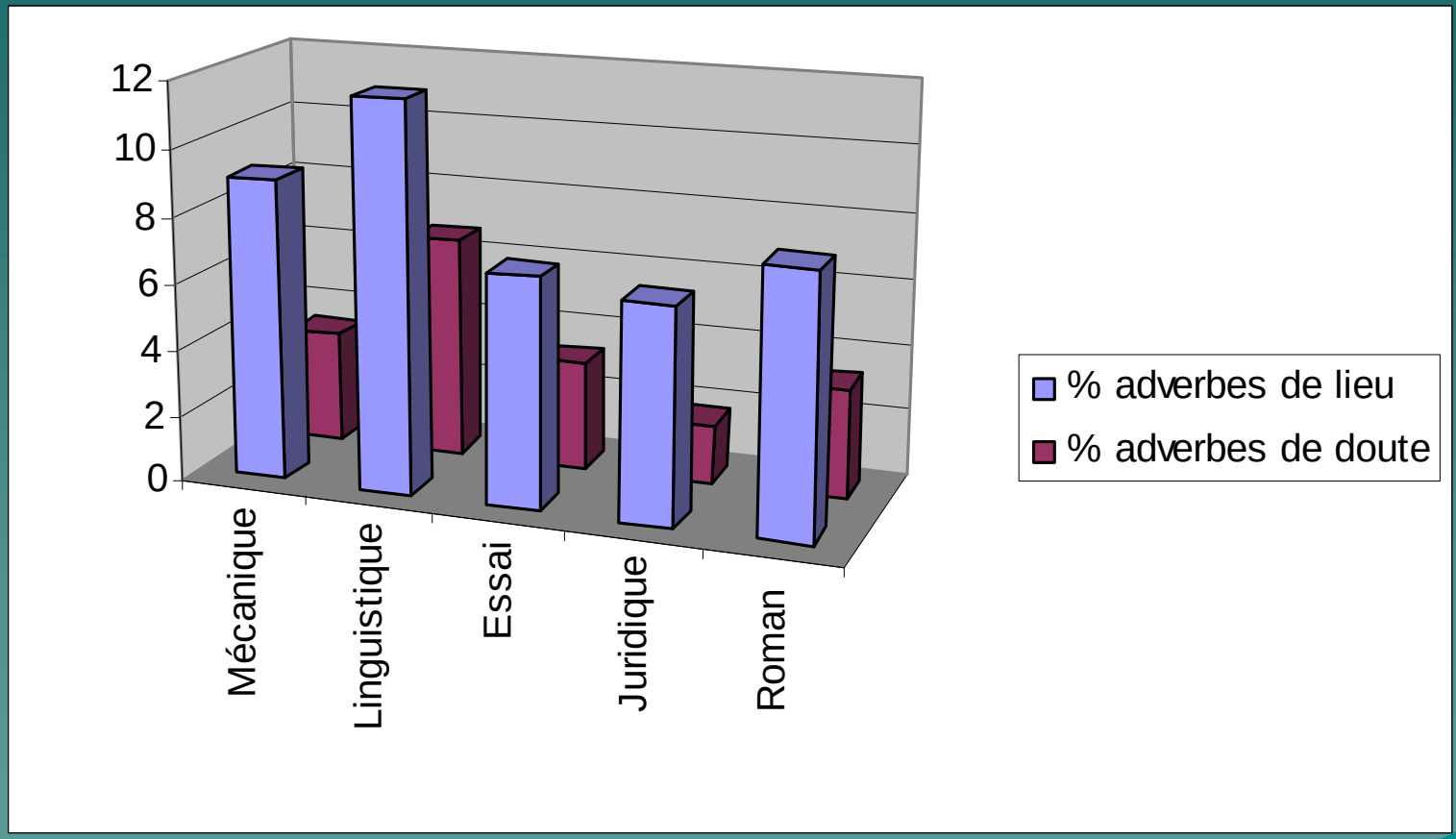
Résultats : mots inconnus



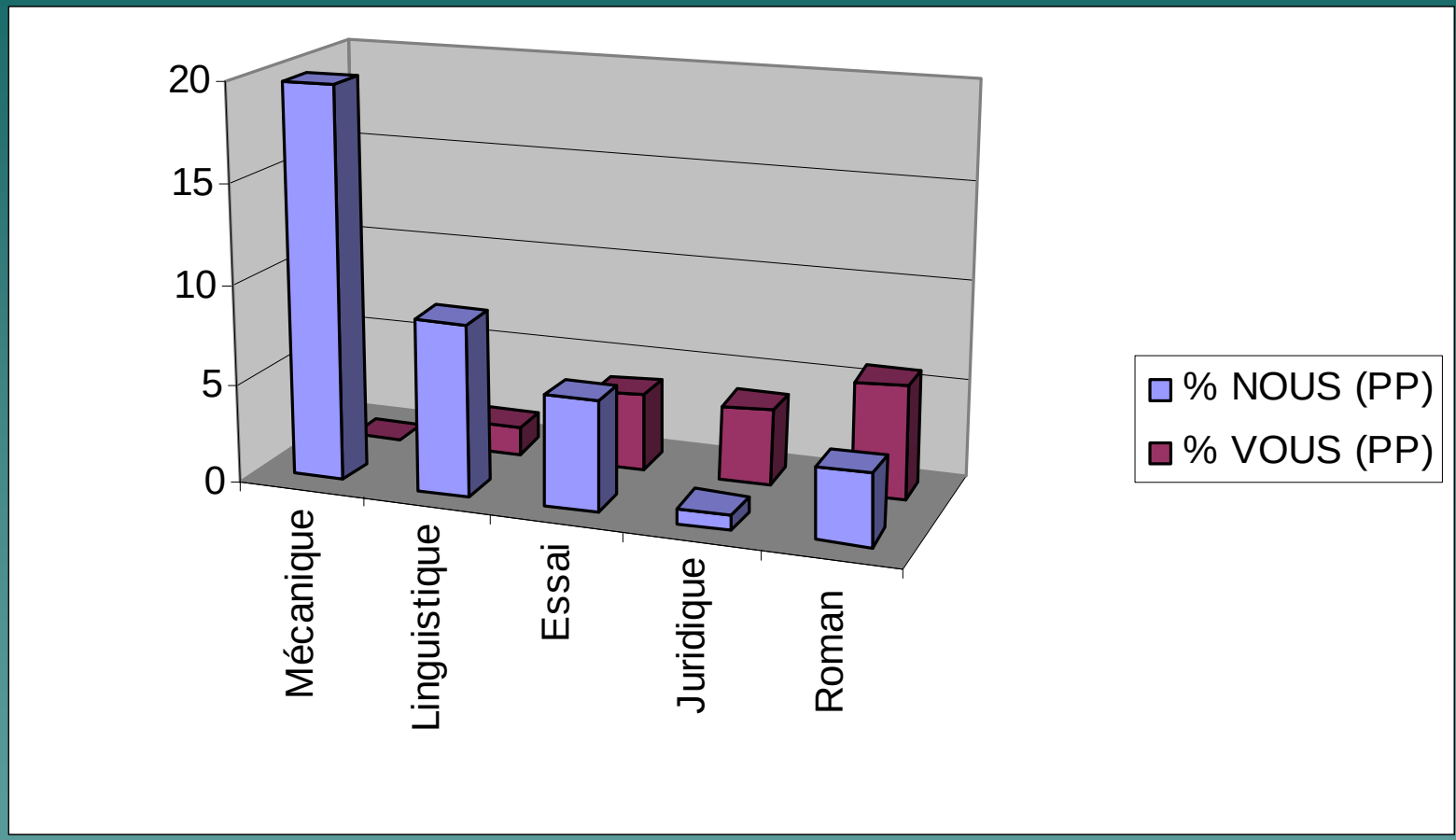
Résultats : abréviations



Résultats : adverbes



Résultats : pronoms personnels



Bilan sur le discours scientifique *versus* autres discours

- La description morphosyntaxique permet de distinguer le discours scientifique des autres discours

- Intérêt de la description pour le genre

Confirmation de ces résultats

- Rinck Fanny, *L'article de recherche en sciences du langage et en lettres : figure de l'auteur et identité disciplinaire du genre*

  Thèse de doctorat soutenue le 17 novembre 2006, Université Stendhal, Grenoble.

Un jeu d'étiquettes pour les textes scientifiques

- Limites de Cordial®

- **Adoption d'un nouveau jeu d'étiquettes**
  - **Étiquettes dédiées aux genre de l'article scientifique**
  - **Entraînement de TnT sur le corpus de linguistique**

15 catégories – 145 variables parmi lesquelles :

- Formalisation
- Adverbes et connecteurs
- Adjectifs
- Pronoms personnels
- Verbes
- Déterminants
- Noms
- Particules
- Ponctuation
- Subordonnants
- Interjections
- Numéraux
- Préfixes

Caractéristiques intéressantes

- 14 marques de ponctuation
- Les verbes de modalité
- 15 types de connecteurs :
  - addition, concession, conclusion, doute,…
- Pronoms personnels
  - *il* anaphorique *il* impersonnel
- Numéraux
  - numéraux, cardinaux, dates, indices de listes

Analyse différentielle des textes scientifiques

- Examen des variables sur et sous-représentées dans les deux corpus

- Analyse significative dès qu'elle est supérieure ou égale à 0,2 seuil de corrélation obtenu sur le corpus d'entraînement.

Variables sous-représentées

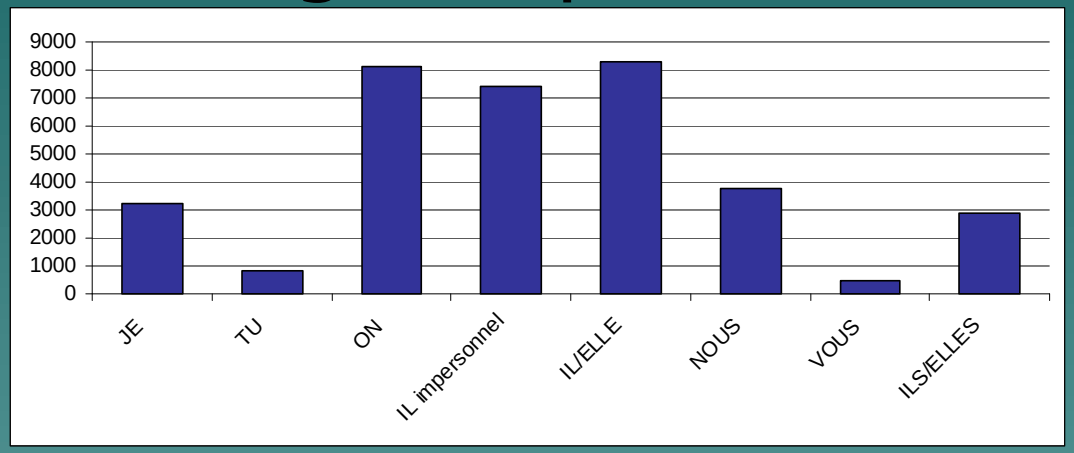
- Absence de marques de 1ère, 2ème personnes du singulier et du pluriel
- Temps peu ou non représentés : subjonctif imparfait, passé simple, conditionnel (sauf verbes de modalité)
- Ponctuation et formalisation mathématique

Les guillemets en mécanique : des effets de style ?

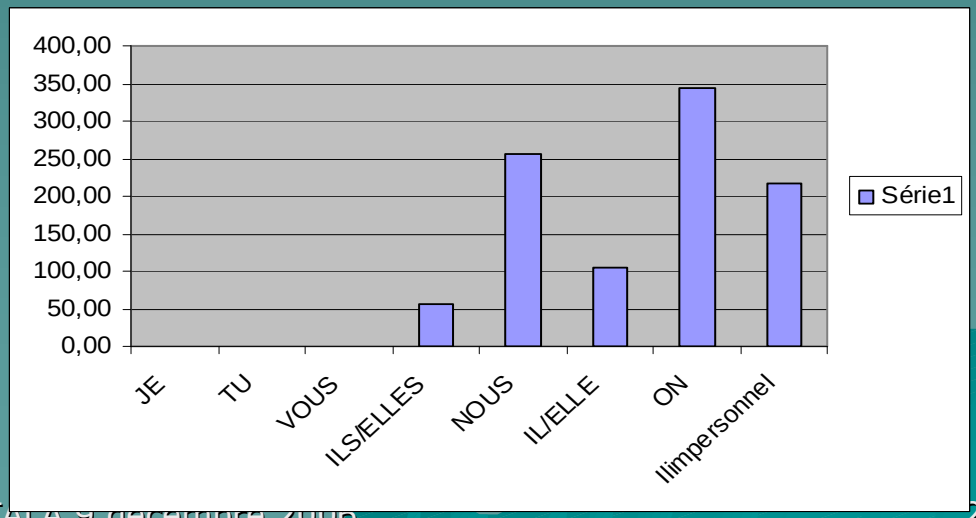
- *« les effets terrifiants du pyroxyle »*
- *« une signature acoustique »,*
- *le repérage de certaines des « Bertha », etc.*

Variables sur-représentées : les personnes

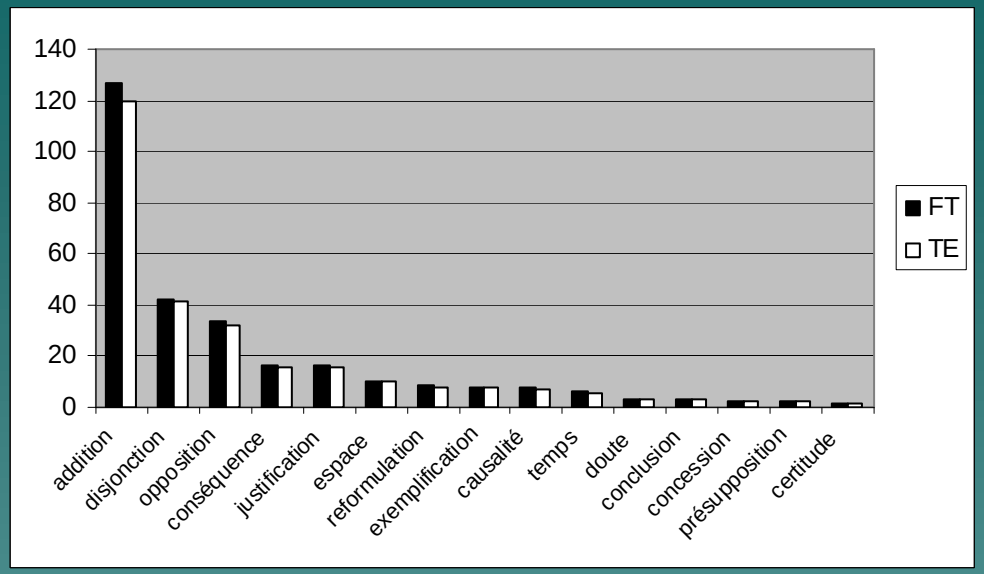
Linguistique



Mécanique

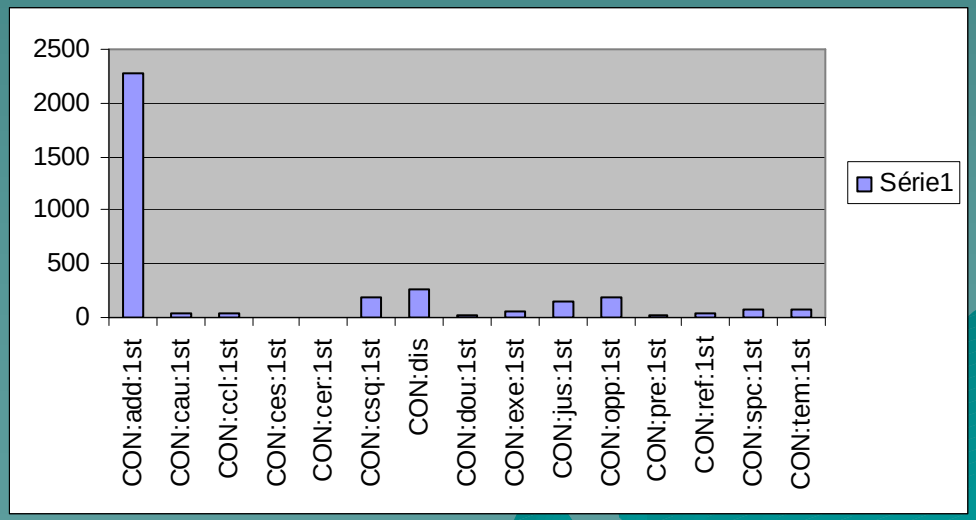


Les variables sur-représentées : les connecteurs

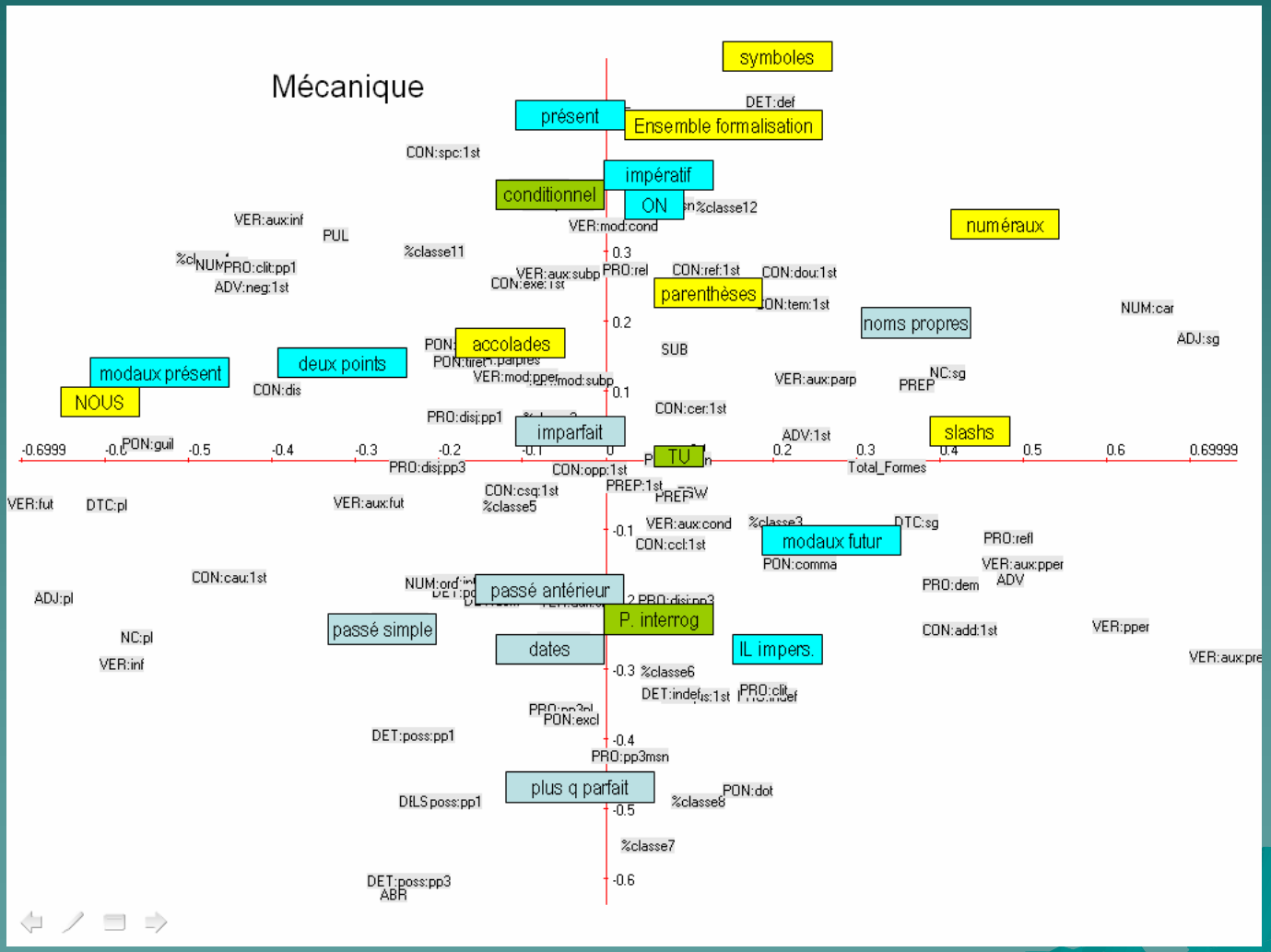


Linguistique

Mécanique



ACP sur les textes de mécanique
(Poudat, 2006)



Bilan de l'étude

- Affiner la description de certaines variables pour la mécanique
  - Morphologie et adjectifs

    atraumatique,

    thermoacoustique

    poroélastique

    vibroacoustique

  - Symboles et ponctuation

Le genre et la recherche d’information

Classification

- Le point de départ de nombreuses applications
  - Analyse lexicale ou thématique
  - Génération de liens hypertextes, *etc.*
  -
- Classification supervisée ou non supervisée
- Classer des textes ou des parties de textes
- Représentations variables du texte

Application à la classification de textes

POUDAT C., CLEUZIOU G. et CLAVIER V. (2006). « Catégorisation de textes en domaines et genres : complémentarité des indexations lexicale et morpho-syntaxique » *in* Document numérique vol. 9, n°1, Paris : Hermès, Editions Lavoisier, pp. 25-42.

Classification en genres et domaines

- Classer des textes
  - En genres
  - En domaines
- Corpus pilote : 371 textes
  - 3 genres : articles, présentations de revues et comptes rendus
  - 2 domaines : linguistique et mécanique
- Descripteurs lexicaux, morphosyntaxiques ou les deux combinés

Méthode de classification

- Guillaume Cleuziou, *Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information*

  Thèse de doctorat, Université d'Orléans, 2004

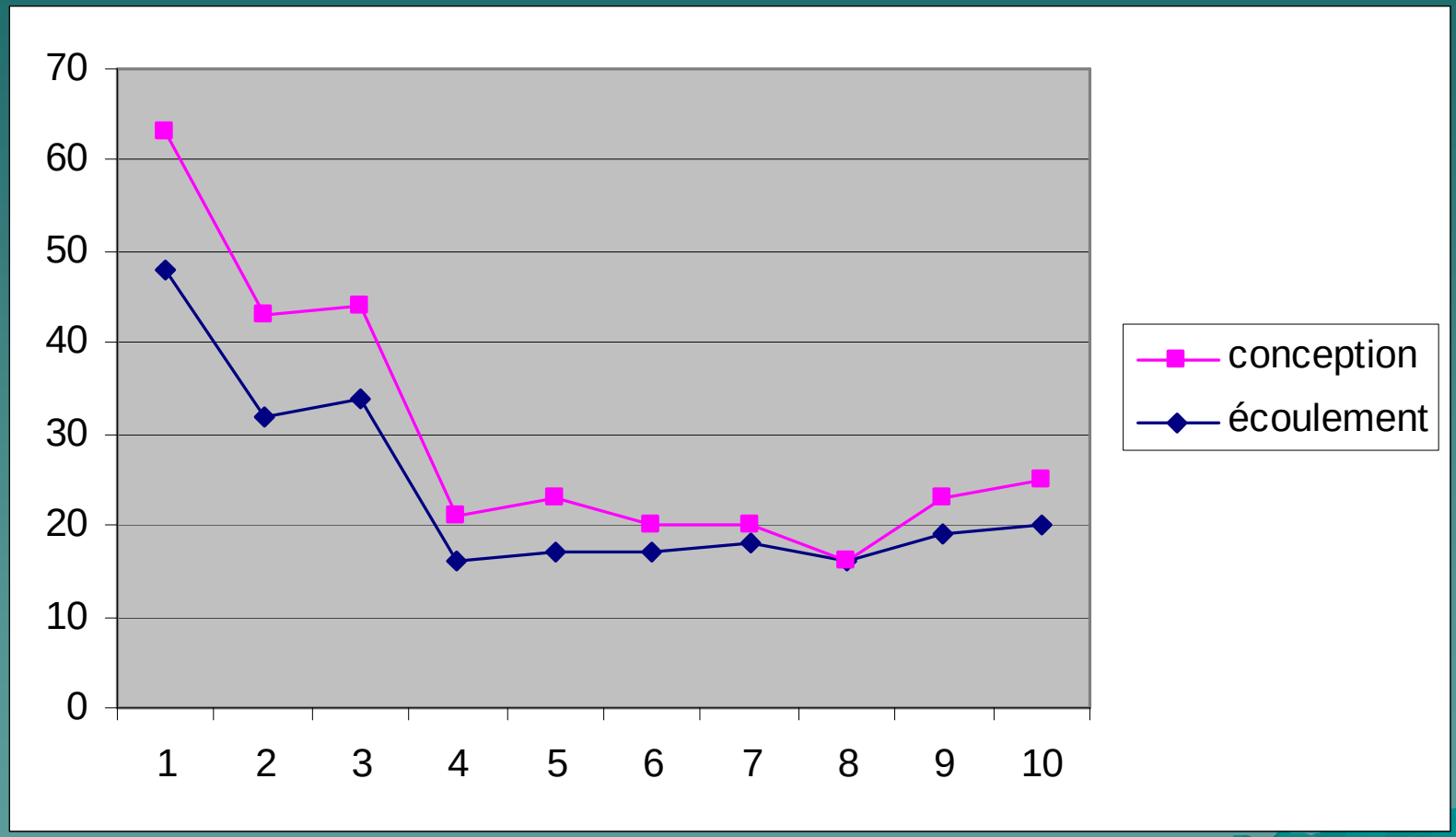
Principaux résultats

- Classification en domaine
  - Les descripteurs morphosyntaxiques sont plus discriminants que les descripteurs lexicaux
  - Meilleurs résultats avec utilisation conjointe des descripteurs
- Classification en genre
  - Même constat (Lee & Myaeng, 2002)

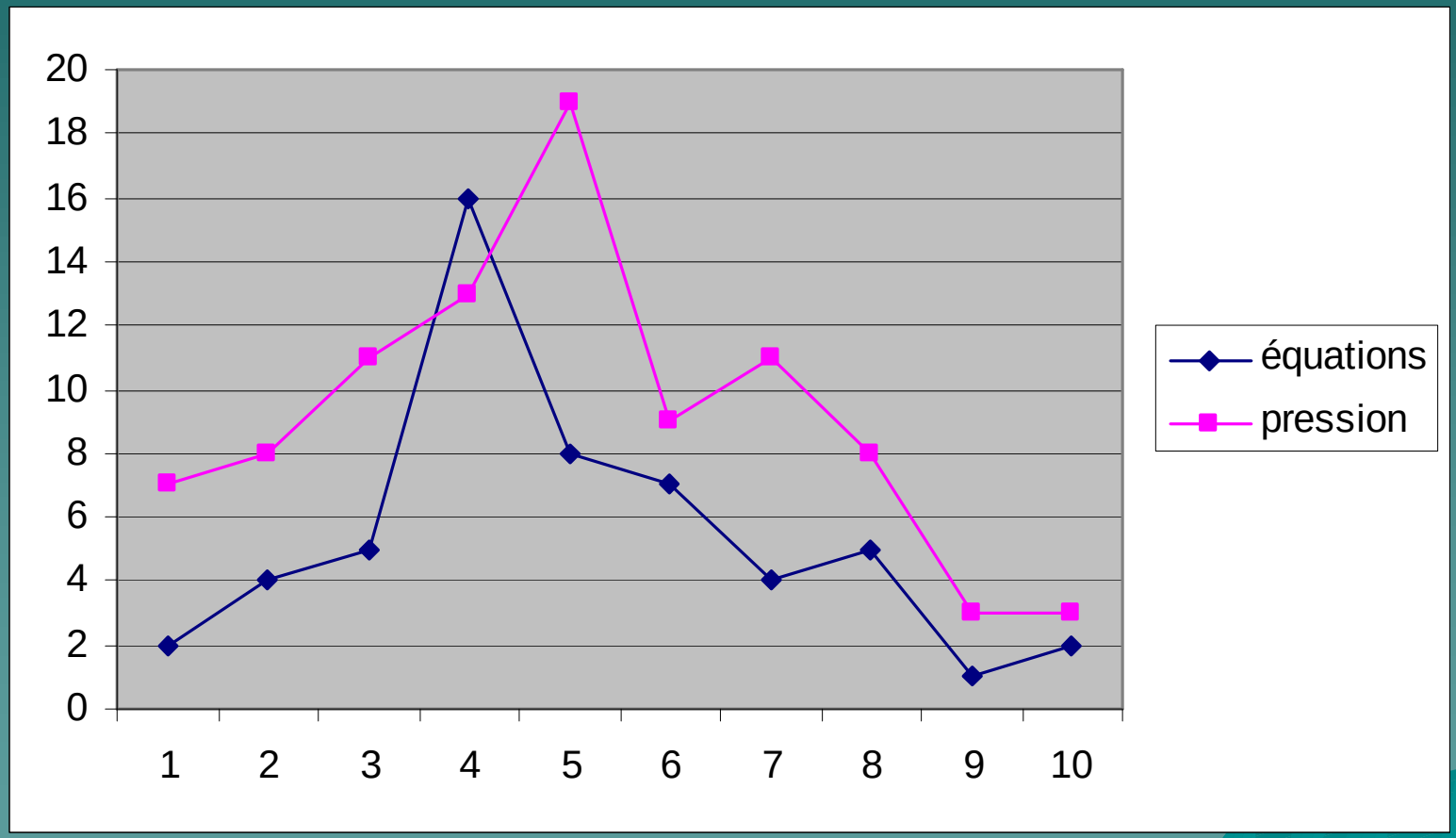
Indexation

- Terminologie, concepts scientifiques
- Le sens des concepts scientifiques en mécanique varie suivant leur position dans les textes

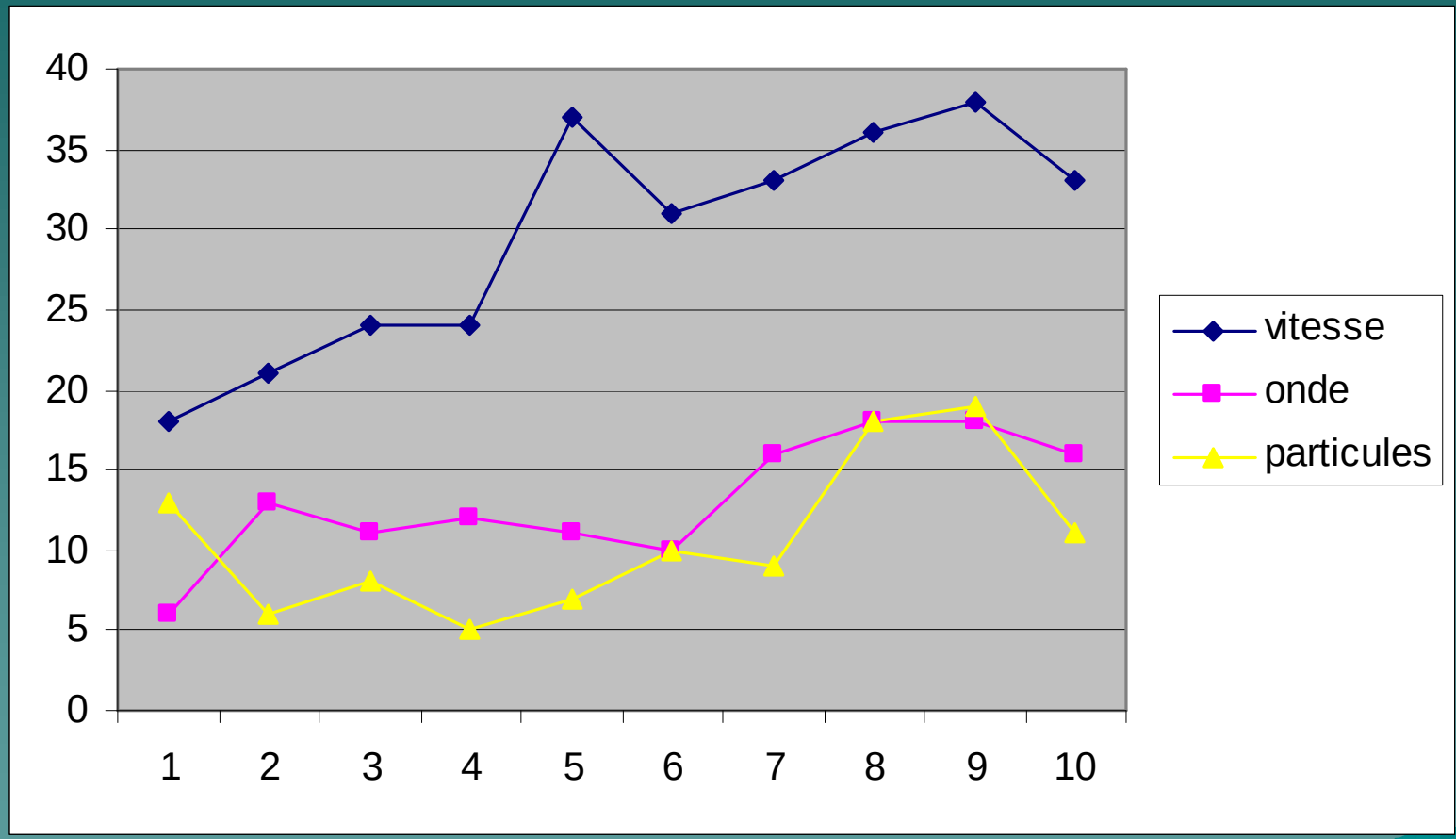
Les concepts de début d'article



Les concepts de milieu d'article



Les concepts de fin d'articles



Indexation et structuration logique

- Indexation et documents structurés
- Structure logique et plan IMRaD en mécanique
- Pas de plan IMRaD dans les textes de linguistique

Conclusion

- Impact de l'identité disciplinaire sur les genres
- Caractéristiques des sciences humaines *versus* des sciences de la nature
  - Positionnement des auteurs dans le champ scientifique, degré d'internationalisation
  - Mode d'exposition de la science
    - La preuve en mécanique : monstration, démonstration, reproductibilité
    - La rhétorique, la polimicité en linguistique