Mise en place d ’une méthodologie d ’analyse automatique de l ’information textuelle sans dépendance de la langue pour la veille technologique

Badreddine Roudani

Plan de l ’exposé

Problématique

Définition de la Veille Technologique (VT)

Définition d ’Analyse Automatique de l ’Information Textuelle (AAIT)

Les différentes méthodes d ’AAIT

Place de notre travail dans le processus d ’AAIT

Notre contribution : système d ’AAIT sans dépendance de la langue

- La lemmatisation
- L ’extraction terminologique
- La consultation directe des documents
- Applications et résultats

- Conclusion et perspectives

Problématique

Nécessité et intérêt
d'une méthodologie d 'analyse automatique
sans dépendance de la langue

⇨ La lemmatisation
- Les approches linguistiques
- Les approches morphologiques

⇨ L'extraction terminologique
- Les approches linguistiques ou statistiques

⇨ La recherche documentaire

La veille Technologique (VT)

Un ensemble de moyens matériels et humains mis en place par une entreprise pour gérer :

- la recherche et la collecte de tous types d ’IST concernant un sujet de surveillance prioritaire,

- le traitement et l ’analyse de l ’information collectée,

- la diffusion de l ’information élaborée pour la prise de décision.

L ’analyse Automatique de l ’information (AAI)

Est l ’application des méthodes mathématiques et statistiques sur un corpus de données bibliographiques et / ou textuelles pour obtenir une vision synthétique, objective et complète de son contenu.

AAI est un outil efficace de VT

Les différentes méthodes d ’AAI

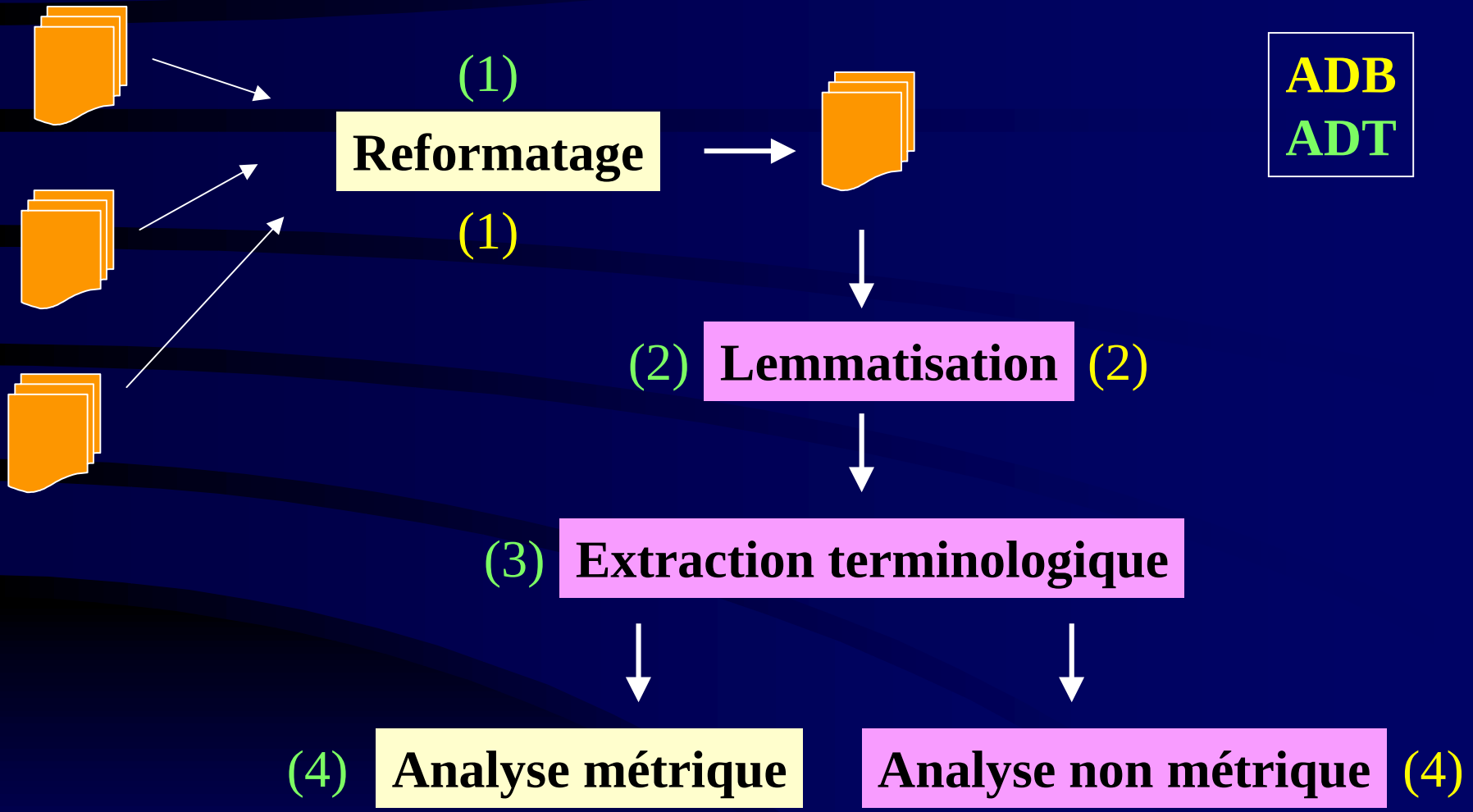
- L ’analyse de données bibliographiques (ADB)

  - La bibliométrie
  - La scientométrie
  - L ’infométrie

- L ’analyse de données textuelles (ADT)

  - La statistique textuelle
  - La statistique linguistique

Place de notre travail dans le processus d ’AAI



La lemmatisation

Est un traitement qui consiste à soumettre les unités d ’analyse à une normalisation qui permet de ramener sous un vocable unique toutes les variantes graphiques d ’un élément d ’analyse

Approche contextuelle

- Technique d ’analyse morpho-syntaxique

Approche hors contexte

- Technique de substitution
- Technique de radicalisation
- Technique de troncature

La lemmatisation par parties communes

Trois possibilités de regroupement :

Préfixe commun : regulate+
regulate
regulated

Suffixe commun : +membrane
transmembrane
trans-membrane

Sous-chaîne commune : +regulat+
regulate
regulated
autoregulation

La lemmatisation par parties communes

Coefficient de similitude

$$C = \text{Long (Partie commune)} / \text{Long (Mot}_i\text{)}$$

Avec $0 < C < 1$

Préfixe commun et C = 0.4

auto-activation
autoactivation
autoregulation
automobile

Sous-chaîne commune et C = 0.8

auto-activation
autoactivation

La lemmatisation par parties communes

Avec cette technique on peut regrouper :

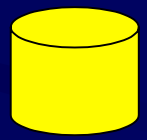
- toutes les formes conjuguées de chaque verbe
- tous les verbes et leur nominalisation
- toutes les formes plurielles et singulières de chaque mot
- toutes les variations graphiques des termes mal orthographiés

DATALEM

Corpus initial

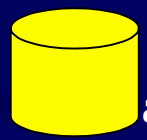
Regroupement par parties communes

Edition pour validation



Lexique terminologique

Substitution



Lexique en automate à états finis

Corpus normalisé

L ’extraction terminologique

C ’est le passage d ’une représentation du texte intégral sous la forme d ’une succession de mots isolés à une représentation sous forme d ’une suite d ’unités terminologiques.

Analyse syntaxique

Segments répétés

Cooccurrence de mots

L ’extraction par multiformes communes

➪ **Une multiforme commune** est un ensemble de mots communs à un groupe d ’éléments textuels.

➪ **Un élément textuel** est une succession de mots délimitée par un caractère de ponctuation.

X A X X B X X C X X X X X X X X D X
X X X X X B X A X X X C X X X X X X X X X D X X X X X

Formes communes : <A> < B> <C > <D>
Eléments textuels :
X A X X B X X C X X X X X X X X D X
X X X X X B X A X X X C X X X X X X X X X D X X X X X

L ’extraction par multiformes communes

Objectif : identifier toutes les structures syntaxiques des unités terminologiques répétées

Formes communes : <synthèse> <loi> <commande>

Eléments textuels :

synthèse d ’une loi de commande
synthèse de loi de commande
synthèse de la loi de commande

Système de réglage :

1- Mots vides
2- Intervalle de voisinage
3- Nombre de formes communes
4- Regroupement avec ordre
5- Regroupement sans ordre

L ’extraction par multiformes communes

Regroupement avec paramétrage

X X B X A X X C X X X X X X D X
X X X A X X B X C X X X X X X X D X X X
X X X A B C X X X X X X D X
X X X X X A X X B X C X X X

I = 14 / N = 4 / sans ordre

Formes communes : <A> < B> <C > <D>
Eléments textuels :
B X A X X C X X X X X X D
A X X B X C X X X X X X X D
A B C X X X X X X D
A X X B X C

I = 6 / N = 3 / sans ordre

Formes communes : <A> < B> <C >
Eléments textuels :
B X A X X C
A X X B X C
A B C
A X X B X C

Formes communes : <energy> <surface>
Eléments textuels :

energy of si surface
energy dissipation in sliding crystal surface
energy of a nember of surface
energy necessary to archive a given surface
surface state energy

Formes communes : <electron> <collision>
Eléments textuels :

collision strengths for electron
electron and hole collision
electron atom ionizing collision
electron h2 collision
electron molecule collision

Formes communes : <external> <field>
Eléments textuels :

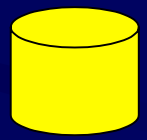
external bias field
external electric field
external magnitic field

DATALEX

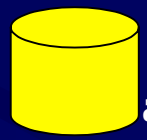
Corpus de textes lemmatisés

Regroupement par multiformes communes

❸ Edition pour validation



Lexique d ’unités terminologique



Lexique en automate à états finis

❹ Indexation

Corpus de textes transformés en unités terminologiques

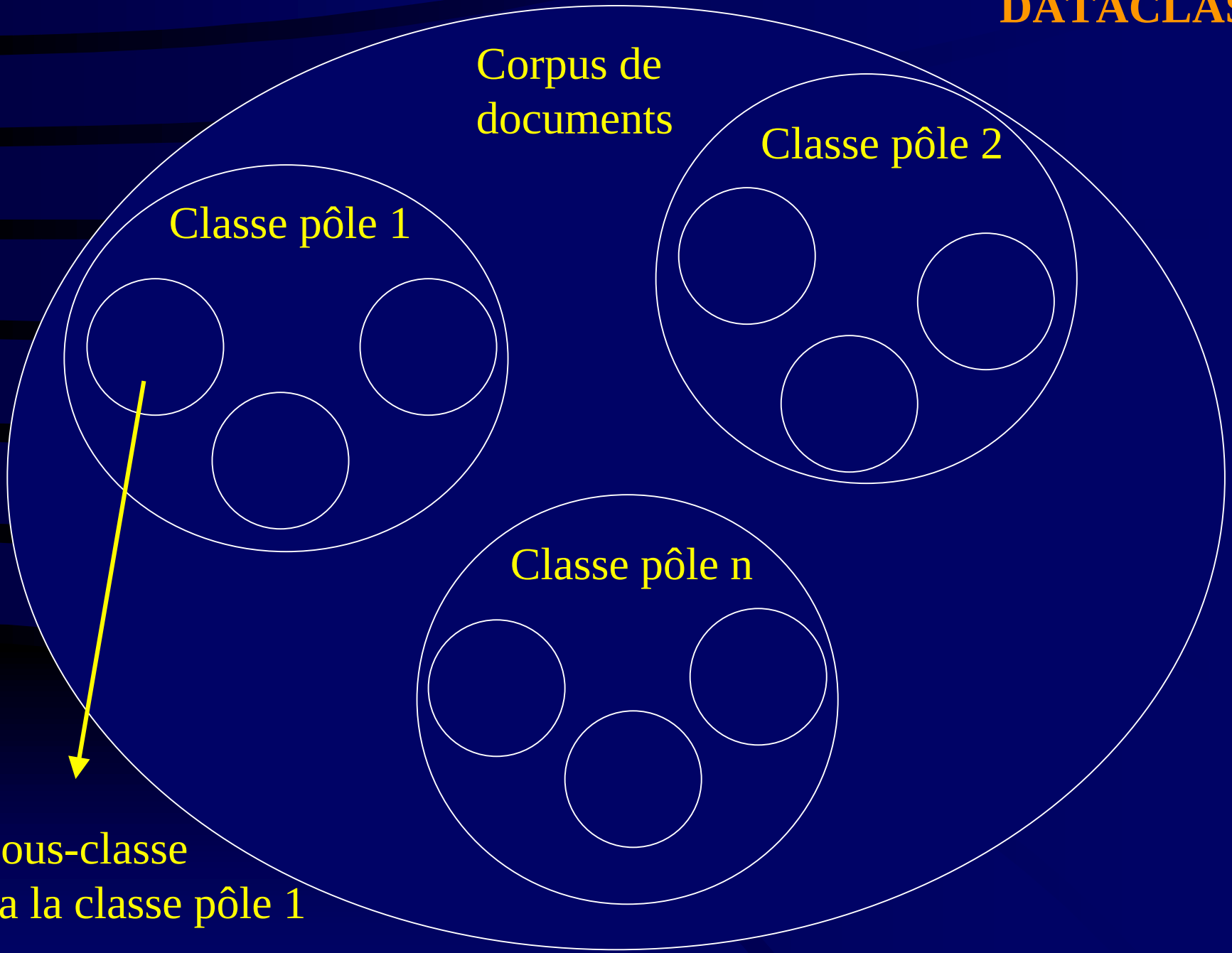
La consultation directe des documents

Les outils existants :

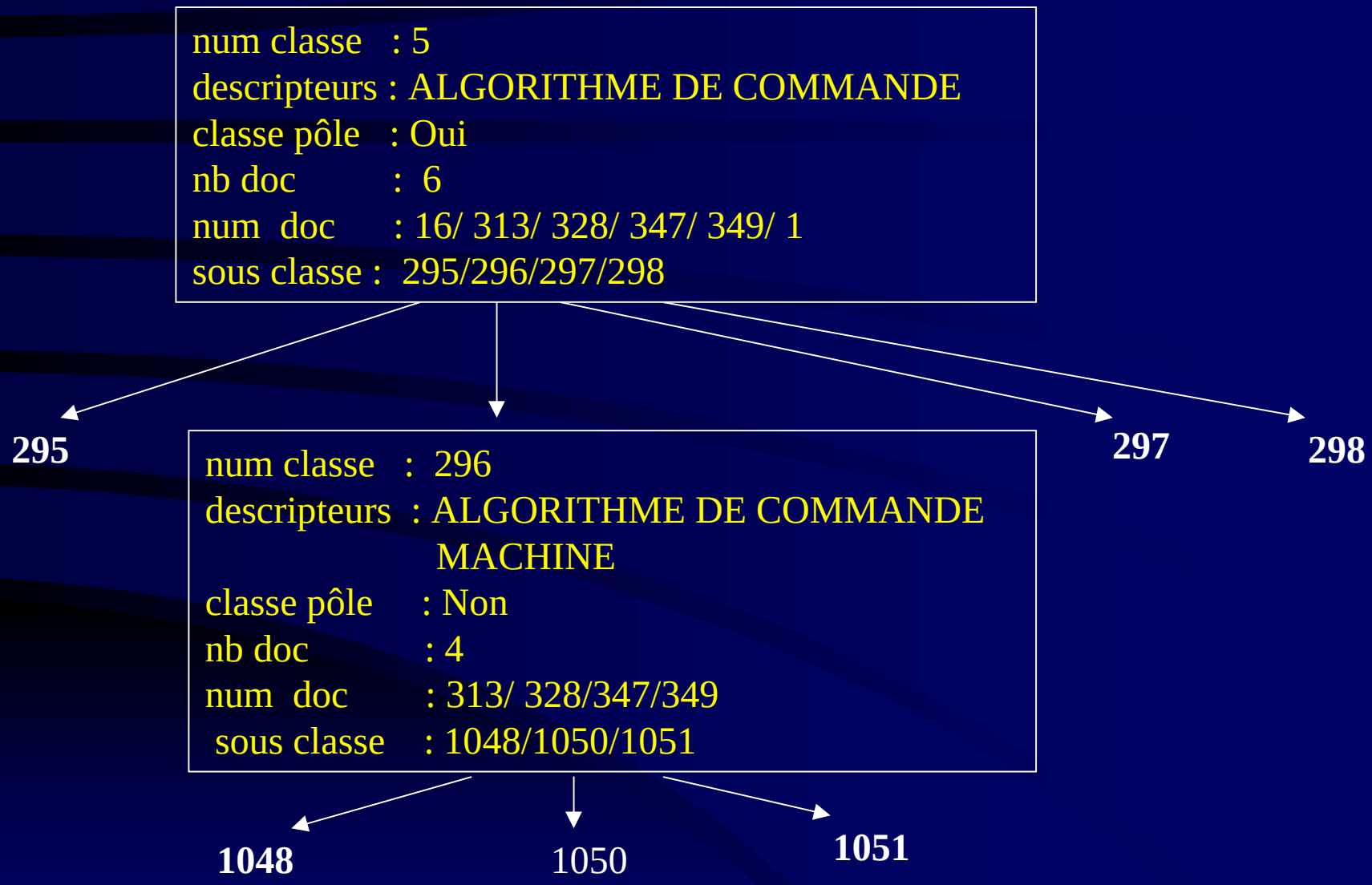
- Les systèmes classiques de GED
- Les systèmes d ’interrogation des BD en langage naturel
- Les systèmes de navigation hypertexte

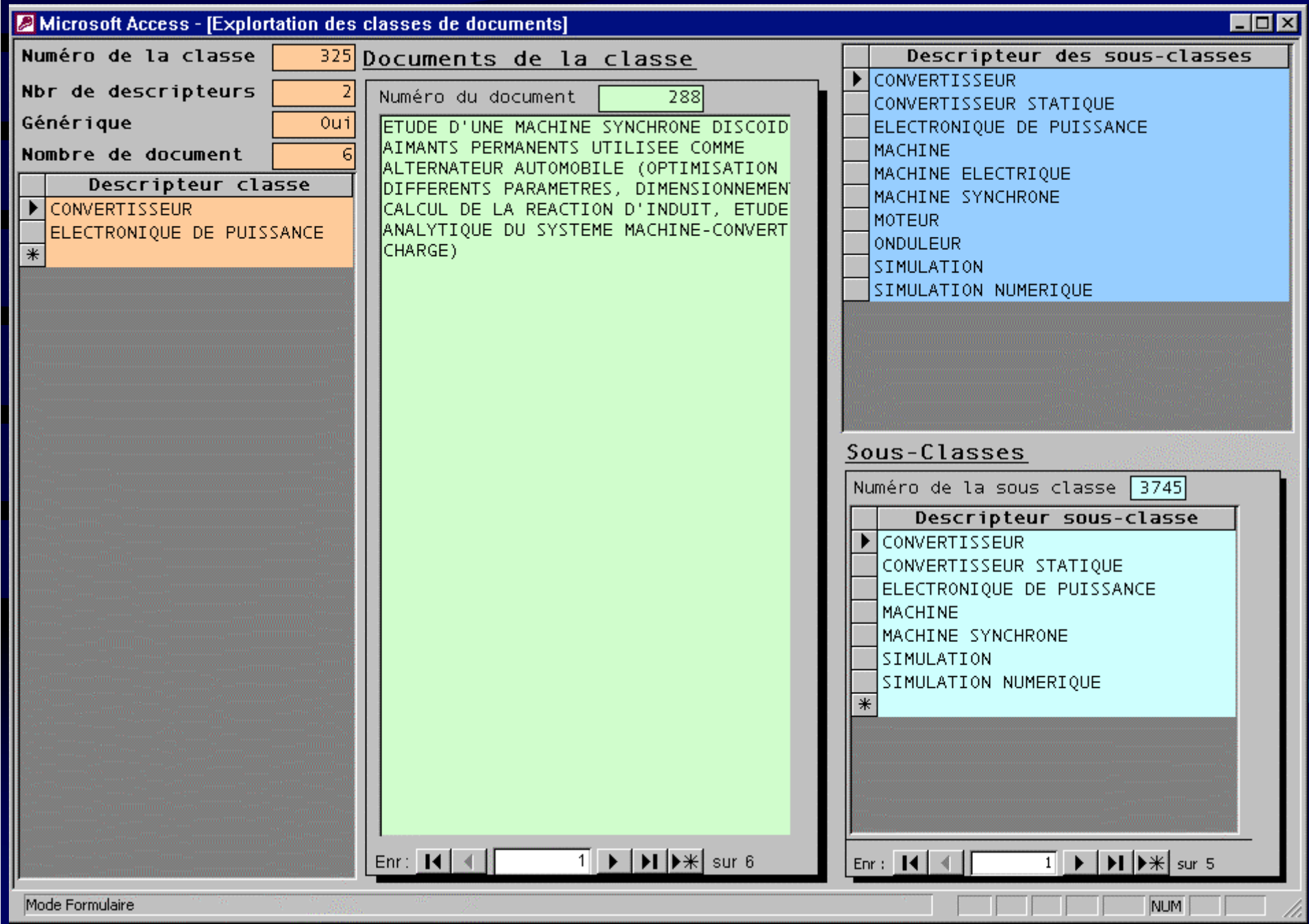
Notre outil : système de classement hiérarchique non métrique

La répartition hiérarchique	DATACLASS
La consultation et la navigation	ACCESS


Corpus de documents
Classe pôle 1
Classe pôle 2
Classe pôle n
Sous-classe da la classe pôle 1

Arbre de classement hiérarchique : DATACLASS

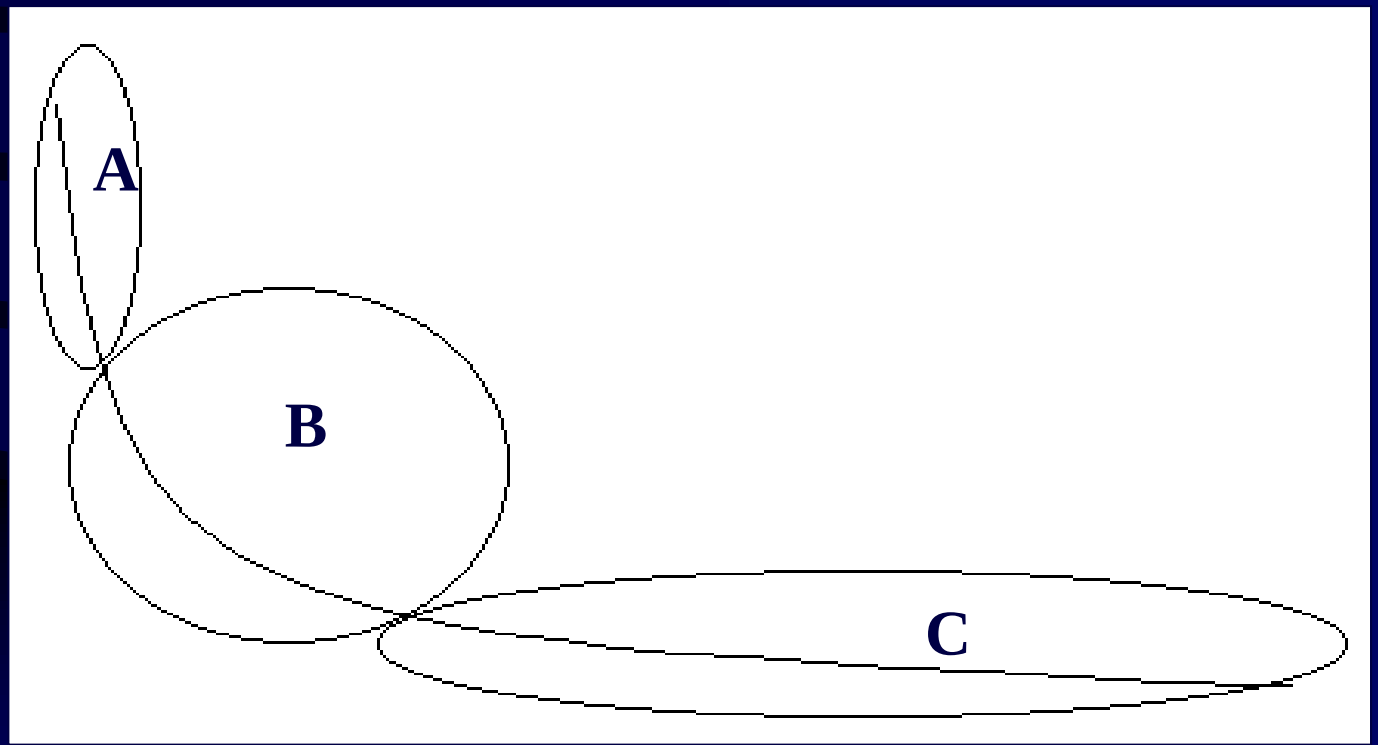



Microsoft Access - [Explortation des classes de documents]
Numéro de la classe
325
Nbr de descripteurs
2
Générique
Oui
Nombre de document
6
Descripteur classe
CONVERTISSEUR
ELECTRONIQUE DE PUISSANCE
Documents de la classe
Numéro du document
288
ETUDE D'UNE MACHINE SYNCHRONE DISCOID
AIMANTS PERMANENTS UTILISEE COMME
ALTERNATEUR AUTOMOBILE (OPTIMISATION
DIFFERENTS PARAMETRES, DIMENSIONNEMEN
CALCUL DE LA REACTION D'INDUIT, ETUDE
ANALYTIQUE DU SYSTEME MACHINE-CONVERT
CHARGE)
Enr :
1
sur 6
Descripteur des sous-classes
CONVERTISSEUR
CONVERTISSEUR STATIQUE
ELECTRONIQUE DE PUISSANCE
MACHINE
MACHINE ELECTRIQUE
MACHINE SYNCHRONE
MOTEUR
ONDULEUR
SIMULATION
SIMULATION NUMERIQUE
Sous-Classes
Numéro de la sous classe
3745
Descripteur sous-classe
CONVERTISSEUR
CONVERTISSEUR STATIQUE
ELECTRONIQUE DE PUISSANCE
MACHINE
MACHINE SYNCHRONE
SIMULATION
SIMULATION NUMERIQUE
Enr :
1
sur 5
Mode Formulaire
NUM

Application 1 : la lemmatisation

Objectif

réduire la dispersion du vocabulaire utilisé.



Résultats

Mots clés non contrôlés : DOC-THESE

Indicateurs statistique	Avant lemmatisation	Après lemmatisation
Nombre total de formes	1946	1662
Nombre de formes (fréquence < 3)	1714	1400

Termes du texte libre : FLYBASE

Indicateurs stat	Nombre total de formes	Nb de formes à fré <=7
Vocabulaire brut	2582	2237
Lexique complété	1967	1630
DATALEM	1945	1595

Application 2 : Le redressement des affiliations

Objectif

mettre en évidence les variations d ’écriture des affiliations

PASCAL 4702 réf :

4702 —Dedoublonnage→ 3470

3470 —Tri alphabétique→ 2291

3470 —Multiformes communes→ 1585

Exemple de regroupement

Tri alphabétique :

CNRS, cent. physique theorique Marseille/13288 Marseille/FRA
CNRS, cent. physique theorique/13288 Marseille/FRA

Multiformes communes :

CNRS, cent. physique theorique Marseille/13288 Marseille/FRA
CNRS, cent. physique theorique/13288 Marseille/FRA
C.N.R.S., cent. physique theorique Marseille/13288 Marseille/FRA
det physique theorique/CNRS/13288 Marseille/FRA
cent. physique theorique / Marseille 13288 /FRA

Application 3 : L ’extraction terminologique

Objectif

Indexation d ’un corpus de résumés

Matériel de base : 497 résumés / DOC-THESE

Démarche :

1- Copie du corpus initial
2- Segmentation
3- Lemmatisation
4- Détection des unités terminologiques et leurs variantes graphiques
5- Indexation

DATALEM

Lexique de 2394 entrées

Forme développée	Forme canonique
ABAISSENT	ABAISSER
ABSORBANTS	ABSORBER
ABSORBANTS	ABSORBER
ABSORBENT	ABSORBER
ABSORBEES	ABSORBER
ABSORBEE	ABSORBER

DATALEX

Nombre de formes communes : 4
Nature de regroupement : sans ordre
intervalle de voisinage : 8

→ 26 termes composés de 4 mots

Nombre de formes communes : 3
Nature de regroupement : sans ordre
intervalle de voisinage : 6

→ 259 termes composés de 3 mots

Nombre de formes communes : 2
Nature de regroupement : sans ordre
intervalle de voisinage : 4

→ 981 termes composés de 2 mots

→ 242 termes simples

Résultat : Lexique de 2394 entrées

Exemple comparatif

Premier résumé

MOTS-CLES D'AUTEUR : SYNTHESE COMMANDE/ COMMANDE ROBUSTE/ CONTRAINTE/ STABILISATION/ ALGORITHME/ OPTIMISATION/ RESEAU ELECTRIQUE/ COUT GARANTI/ INEGALITE MATRICIELLE LINEAIRE / CD

RÉSUMÉ INDEXÉ : SYSTEMES SOUS CONTRAINTE STRUCTURELLE/ SYNTHESE DE LOI DE COMMANDE / COMMANDE ROBUSTE / LOIS DE COMMANDE / ALGORITHME DE COMMANDE / DISTRIBUTION D'ENERGIE ELECTRIQUE / COMMANDE A FAIBLE COUT / SYSTEME DE PUISSANCE / GENERATION / RESOLUTION

Application 4 : La reconnaissance du vocabulaire spécifique

P0 - 107.1 acts in trans to disrupt accumulation of maternal tub transcript
P0 - tra and dsx control early inductive signals that determine the sex of XX germ cells
P0 - Pattern of hh expression in ptc mutants studied

Après traitement

Y - ACT TRANS DISRUPT ACCUMULATE MOTHER TRANSCRIBING
N - CONTROL EARLY INDUCTIVE SIGNAL DETERMINE SEX XX GERM CELL
I - PATTERN EXPRESS MUTANT STUDY

Application 4 : La reconnaissance du vocabulaire spécifique

But :

mettre en évidence l ’ existence de combinaisons de termes qui se retrouvent spécifiquement dans des groupes de phrases qui décrivent l ’interaction.

1200 phrases → DATALEX → 6307 groupes

Exemple de regroupement

FCOM : <DOWNSTREAM> <ACT>

Y - ACT DOWNSTREAM DER-PATHWAY
Y - ACT DOWNSTREAM
Y - DOUBLE MUTANT ANALYSE SUGGEST ACT DOWNSTREAM ACTIVATE
Y - GENE PRODUCT ACT SELECT HOMEODOMAIN PROTEIN INCLUDE DNA BIND TRANSCRIBING FACTOR ALTER REGULATE DOWNSTREAM TARGET GENE
Y - ACT HOMEOTIC GENE DOWNSTREAM TERMINAL GAP GENE PROMOTE MORPHOGENETIC DIFFERENTIAL ANTERIOR POSTERIOR MIDGUT
Y - GENE PRODUCT ACT SELECT HOMEODOMAIN PROTEIN INCLUDE DNA BIND TRANSCRIBING FACTOR ALTER REGULATE DOWNSTREAM TARGET GENE
Y - ACT DOWNSTREAM MALPIGHIAN TUBULE REGULATE PATH
Y - ACT AFTER INITIAL MESODERM INDUCTION DOWNSTREAM SPECIFIED PRIMORDIAL CELL FATE HEART VISCERA MUSCLE
Y - HOMOZYGOTE FEMALE FEMALE GENOTYPE df(1)hc244,sxl[m1]/ovo STERILE OVARY DEVOID GERM CELL MUST ACT DOWNSTREAM DIFFER PATH
Y - ACT DOWNSTREAM MALPIGHIAN TUBULE REGULATE PATH

Extrait de requêtes

require & normal & pattern
control & path & active
complex & region & interact
element & signal & product
regressing & gene & activate

568 phrases Y / un total de 653 (87 %)
233 phrases N / un total 491 (47,5 %)

Regroupement avec ordre et adjacence

29 phrases
18 Y et 11 N → DATALEX → 27 structures syntaxiques

Regroupement avec ordre : oui
Regroupement avec adjacence : oui
Ignorer les mots vides : oui
Intervalle de voisinage : 8
Taux de phrases Y : 100 %

Taux % = (Nombre de phrases Y / Nombre total des phrases) x 100

Structure syntaxique : [nomgen] **(4)** interact **(2)** [nomgen]

Phrases :

- **Y** because **[nomgen]** mutant allele exhibit dose-sensitive **interact** with **[nomgen]** lack-of-function mutant we have investigate whether the **[nomgen]** or **[nomgen]** protein direct **interact** with the **[nomgen]** protein in vitro
- **Y** we also finder that most **[nomgen]** allele that do not **interact** with **[nomgen]** are missense mutant in the C-terminal EGF and C1r/s repeat or encode truncate protein that delete these repeat
- **Y** **[nomgen]** mutant also **interact** with the **[nomgen]** mutant but do not **interact** with other type of **[nomgen]** mutant

Conclusion et perspectives