

# ISTEX-R

Yannick Toussaint (LORIA)

Pascal Cuxac (INIST)

# Membres du projet

- ATILF :
  - *Evelyne Jacquy, Laurence Kister, Bertrand Gaiffe, Etienne Petitjean et Sandrine Ollinger*
- LORIA :
  - Equipe ORPAILLEUR: *Yannick Toussaint*
  - Equipe Synalp: *Jean-Charles Lamirel, Christophe Cerisara*
- INIST :
  - Service recherche developpement et experimentation(SRDE) : *Sabine Barreaux, Dominique Besagni, Pascal Cuxac, Claire François, Ivana Roche*

*Coordinateur : Yannick Toussaint (LORIA)*

# Objectifs

- Projet de recherche appliquée
  - Intégrer et à mettre à disposition des outils d'accès au contenu
  - Opérer sur des textes intégraux
  - Construire et capitaliser des connaissances sur des domaines scientifiques ou techniques.
- Produire un démonstrateur

# Objectifs

(Hypothèses de départ)

- l'utilisateur dispose d'un certain nombre d'outils avancés d'accès à l'information (ayant déjà fait leurs preuves pour la recherche d'information)
  - collecte et la consolidation de corpus
  - structurer à faible coût un grand volume de textes
  - Sélectionner un corpus de taille *raisonnable* (quelques centaines, quelques milliers de textes)



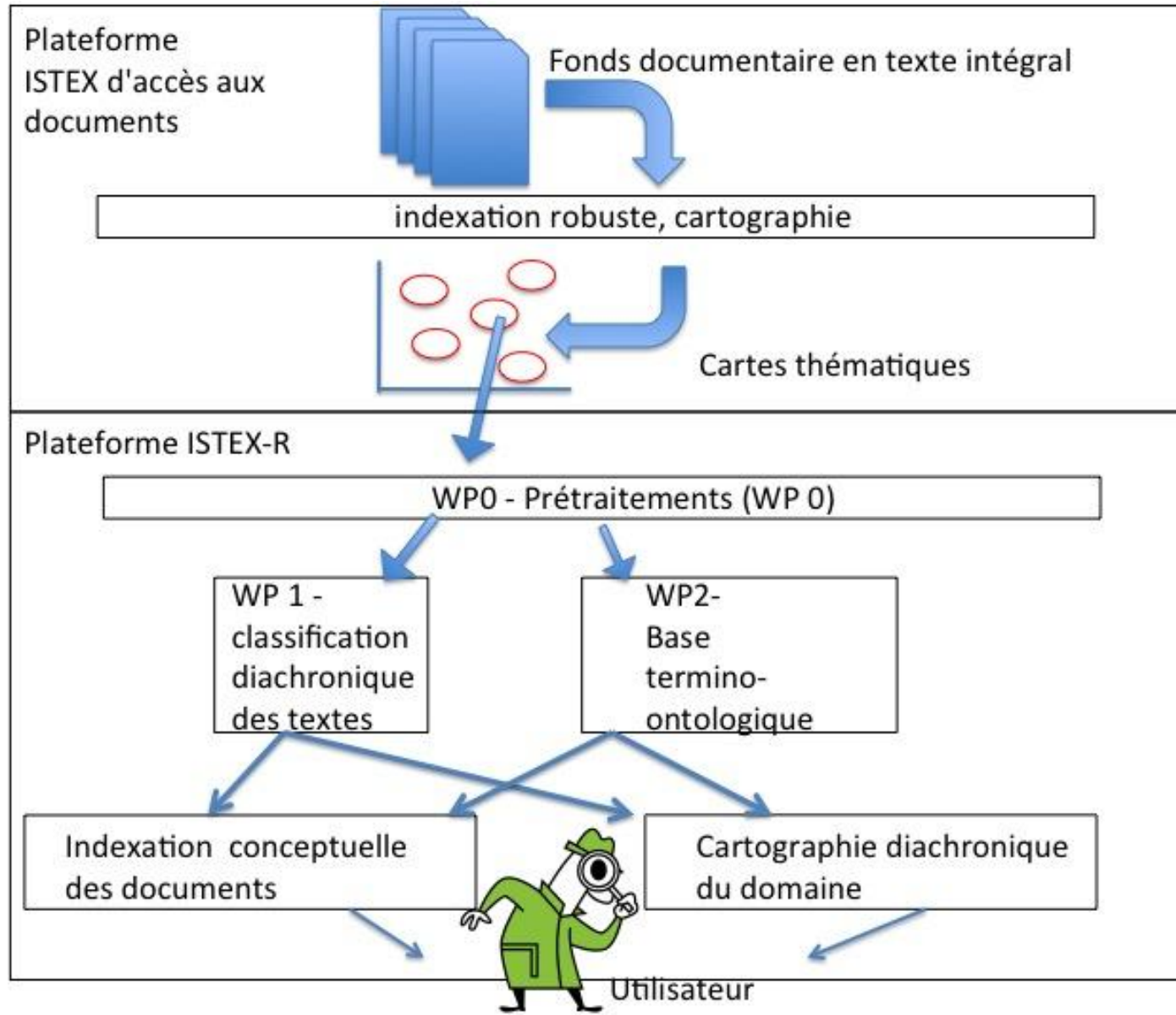
# Objectifs

- vers une analyse plus fine du contenu
  - conceptualisation d'un domaine qui permet de capitaliser les connaissances exprimées au travers des textes
    - mettre en oeuvre une indexation terminologique de qualité
    - Construire des concepts organisée dans une structure similaire à la notion d'ontologie (hiérarchie)
    - Donner les outils pour explorer cette conceptualisation et accéder aux documents impliquant les concepts qu'il recherche.

# Objectifs

- vers une analyse plus fine du contenu
  - comment caractériser l'évolution des recherches et des connaissances dans le temps ?
    - révolution ou électrochoc ?
    - Évolution par des glissements plus subtils d'une problématique vers une autre et par un enrichissement progressif des connaissances
    - La construction de cartes diachroniques vise à outiller l'expert d'un domaine pour lui permettre d'observer ce type d'évolution

# Schéma général du projet



# Prétraitement des textes (WP 0)

- ATILF, LORIA, INIST. Coordinatrice : Evelyne Jacquey
- Représentation des articles en XML - TEI
- Enrichissements linguistiques par annotation des articles :
  - Annotation morpho-syntaxique
  - Annotation syntaxiques
  - Annotation terminologique

# Analyses et cartographies diachronique (WP1)

- LORIA, INIST, Coord. Jean-Charles Lamirel
- Diachronie
  - méthode de classification automatique sur des données associées des périodes de temps successives, et sur l'étude de l'évolution des résultats de classification obtenus
- Clustering incrémental
- Visualisation des résultats des approches incrémentales

# Extraction de connaissances à partir de textes (WP2)

- méthodes issues du traitement automatique de la langue
- de la fouille de données
- de la représentation de connaissances

# Extraction de connaissances à partir de textes (WP2)

- extraction de termes, désambiguïsation et leur utilisation pour l'indexation en texte intégral. (lien avec ANR Termith)
- construction de concepts associés aux termes.
- Etudier la diachronie des termes (lien entre extraction de connaissances et analyses et de cartographies diachroniques).

# Extraction de connaissances versus recherche d'information

- Recherche d'information
  - fournit en réponse à la requête d'un utilisateur, un ensemble de documents les plus appropriés
  - caractérise généralement le document par des mots-clés pondérés
  - cherche essentiellement à regrouper ou à distinguer les documents.
- L'extraction de connaissances
  - s'intéresse au contenu et vise donc à extraire les objets d'un domaine et à les caractériser par des propriétés
  - extrait des relations entre ces objets
  - synthétise le contenu d'un ensemble de textes.
- Les deux approches reposent sur des méthodes de classification
  - la RI classe des documents
  - ql'EC classe des entités du domaine en fonction de leurs caractéristiques.



# Extraction de connaissances versus recherche d'information

- Besoin de synthèse existe dans tous les domaines
  - dans la vie courante
  - enjeu crucial pour les domaines scientifiques ou technologiques
    - pouvoir acquérir ou mettre à jour rapidement ses connaissances sur son domaine d'intérêt
    - lutter contre une trop forte spécialisation liée à des lectures abondantes dans un domaine de plus en plus étroit en donnant accès à des connaissances connexes

*Par exemple, pour un médecin, comprendre les impacts sociologiques d'une maladie.*

# Projection terminologique (metamap)

Fluoroquinolone resistance (FQ-R) in clinical isolates of Enterobacteriaceae species has been reporting frequency in recent years. Two mechanisms of FQ-R have been identified in gram-negatives: mutations in DNA gyrase and reduced intracellular drug accumulation. A single point mutation has been shown to reduce susceptibility to fluoroquinolones. To determine the extent of *gyrA* mutations associated with FQ-R in enteric bacteria, one set of oligonucleotide primers was selected from conserved regions in the flanking regions of the quinolone resistance-determining regions (QRDR) of *Escherichia coli*. This set of primers was used to amplify and sequence the QRDRs of Enterobacteriaceae type strains and 60 fluoroquinolone-resistant clinical isolates of *Citrobacter freundii*, *Enterobacter cloacae*, *E. coli*, *K. pneumoniae*, *Klebsiella oxytoca*, *Providencia stuartii*, and *Providencia marsevensis*. Although similarity of the nucleotide sequences of seven species ranged from 85% to 95% when compared with that of *E. coli*, the amino acid sequences of the *gyrA* QRDR were compared. Conservative amino acid substitutions were detected in the QRDRs of the susceptible type strains: *E. coli* (Ser-83 to Thr), and *P. stuartii* (Asp-87 to Glu). Strains with ciprofloxacin MICs of  $\geq 2$  microg/ml expressed amino acid substitutions primarily at the Gly-81, Ser-83, or Thr-86 positions. Fluoroquinolone MICs varied significantly for strains exhibiting identical *gyrA* mutations, indicating that mutations outside *gyrA* contribute to resistance. The type and position of amino acid alterations varied among these six genera. High-level FQ-R frequently was associated with single *gyrA* mutations in Enterobacteriaceae in this study except *E. coli*.

# Extraction de connaissances versus recherche d'information

- Un exemple concret :
  - bilan des connaissances sur la *dystrophie musculaire de Duchenne*
  - élaboré manuellement après une lecture systématique d'environ 150 articles scientifiques alors que plus de 4000 articles sont répertoriés dans PubMed sur cette maladie
- Un problème majeur :
  - Les textes ne sont pas autonomes (connaissances implicites)
  - Il faut pouvoir introduire des connaissances dans le processus de fouille de données

# Extraction de connaissances versus recherche d'information

- Un exemple concret : les maladies rares
  - bilan des connaissances sur la *dystrophie musculaire de Duchenne*
  - élaboré manuellement après une lecture systématique d'environ 150 articles scientifiques alors que plus de 4000 articles sont répertoriés dans PubMed sur cette maladie.

# L'analyse de résumés sur les Maladies Rares

US National Library of Medicine  
National Institutes of Health

Advanced

Help

Abstract ▾

Send to: ▾

Zh Vopr Neurokhir Im N N Burdenko, 2014;78(5):3-15; discussion 15.

## [Surgical management of patients with pathological deformities of carotid arteries].

[Article in Russian]

Usachev DY, Lukshin VA, Sosnin AD, Shishkina LV, Shimgel'skii AV, Nagorskaya IA, Vasil'chenko VV, Belyaev AY, Akhmedov AD, Batishcheva EV.

### Abstract

Surgical management of pathological deformities of the internal carotid arteries, a cause of chronic brain ischemia, is discussed. This pathology is very common and is found in 25% of all individuals who underwent preventive medical examination according to the ultrasonography data. Most deformities do not pose any threat to patients, while some of them may cause ischemic stroke and chronic brain ischemia. The study included 165 patients with the known follow-up history who had been operated on at the N.N. Burdenko Neurosurgical Institute since 2001. A total of 196 reconstructive interventions of carotid arteries were analyzed. The indications for surgical management of pathological deformities based on clinical symptoms and identification of the signs of vascular wall dysplasia are thoroughly discussed. The local and cerebral hemodynamics during pre- and postoperative period are analyzed. The results of pathomorphological examination of the resected fragments of the deformed arteries are presented; they show that the changes are identical to those in patients with fibromuscular dysplasia. The follow-up history of the patients was recorded; it showed a sustained regression of transitory ischemic strokes and cerebral symptoms in most cases (69%). For proper indications for surgical management, reconstructive surgical interventions are a reliable and effective method for treating chronic brain ischemia and preventing recurrent ischemic strokes in patients with deformities of carotid arteries.

PMID: 25406903 [PubMed - indexed for MEDLINE] [Free full text](#)



Publication Types, MeSH Terms ▾

LinkOut - more resources ▾

PubMed Commons

0 comments

[PubMed Commons home](#)

[How to join PubMed Commons](#)

### Full text links



### Save items

☆ Add to Favorites ▾

### Related citations in PubMed

**Review** Fibromuscular dysplasia of the internal carotid artery. Personal ex[Acta Chir Belg. 1999]

Fibromuscular dysplasia of the internal carotid arteries. Clinical experience and [Ann Surg. 1981]

Clinical manifestations and diagnosis of pathological deformity c [Angiol Sosud Khir. 2011]

Occlusive fibromuscular disease of arteries supplying the brain: results [Ann Vasc Surg. 1997]

**Review** [Fibromuscular dysplasia at the internal carotid origin: a case of [No Shinkei Geka. 1993]

[See reviews...](#)

[See all...](#)

### Recent Activity

[Turn Off](#) [Clear](#)

[Surgical management of patients with pathological deformities of carotid arte PubMed

fibromuscular dysplasia[majr] AND



## **Definition**

La dystrophie musculaire de Duchenne (DMD) est une maladie neuromusculaire caractérisée par une atrophie et une faiblesse musculaires progressives dues à une dégénérescence des muscles squelettiques, lisses et cardiaques.

## **Epidemiologie**

La DMD affecte principalement les garçons avec une incidence à la naissance de 1/3 300 garçons. Les filles sont habituellement asymptomatiques mais un faible pourcentage de femmes conductrices présente des formes modérées de la maladie (Forme symptomatique de la dystrophie musculaire de Duchenne et Becker de la femme conductrice ; voir ce terme). La maladie débute chez les garçons pendant l'enfance avec un retard du développement moteur et du développement global.

## **Description clinique**

En général, les garçons atteints de DMD ne réussissent pas à courir ou sauter. La maladie progresse rapidement et l'enfant développe une marche dandinante avec hypertrophie des mollets (signe de Gowers positif). Monter des escaliers devient difficile et l'enfant tombe fréquemment. La marche devient impossible entre 6 et 13 ans, la moyenne étant de 9,5 ans chez les patients non traités par des stéroïdes. Une cardiomyopathie et une insuffisance respiratoire restrictive peuvent entraîner le décès pendant l'adolescence.

## **Etiologie**

La DMD, d'hérédité récessive liée à l'X, est due à des mutations du gène DMD (Xp21.2) qui résultent en un déficit complet en dystrophine, une protéine sub-sarcolémique.

## **Diagnostic**

Le diagnostic se base sur le tableau clinique, les antécédents familiaux et les résultats de laboratoire (taux de créatinine-kinase sérique 100-200 fois plus élevé que la normale). La biopsie musculaire montre une dystrophie et une absence totale de dystrophine. L'analyse moléculaire montre le plus fréquemment des délétions frame-shift (décalage du cadre de lecture), des duplications ou des mutations faux-sens du gène DMD. (...)

## **Traitement**

Une prise en charge pluridisciplinaire est essentielle. La kinésithérapie basée sur les étirements passifs et des orthèses cruro-pédieuses nocturnes ont pour but de réduire les contractures du tendon d'Achille. Un traitement aux corticostéroïdes (prednisolone, prednisone ou deflazacort) est nécessaire. Les corticostéroïdes doivent être administrés au moment où le développement moteur de l'enfant commence à ralentir, ce qui correspond généralement à l'âge de 5-7 ans. Les complications dues à l'utilisation de stéroïdes doivent être prises en charge et incluent la prise en charge du surpoids, (...)

## **Pronostic**

La DMD a un pronostic sévère et l'espérance de vie est significativement réduite avec un décès survenant tôt à l'âge adulte.

# Travaux réalisés

- Problème d'accès à un corpus ISTEEX, travail sur des corpus déjà disponibles (Termith, Hybride)
- Corpus sur le vieillissement (décembre 2014)
  - Intérêt de plusieurs communautés scientifiques sur Nancy (pluri-domaine, sur plusieurs années et plusieurs éditeurs)
  - 7434 articles avec abstracts provenant de 15 revues
- Chantier thématique d'usage (Fouille de textes)...

# Corpus

Année	nbre notices	editeur
1995	36	elsevier
1996	47	elsevier
1997	435	elsevier
1998	678	elsevier
1999	650	elsevier
2000	1007	elsevier(762) + OUP(245)
2001	1157	elsevier(832)+OUP(325)
2002	329	OUP
2003	310	OUP
2004	292	OUP
2005	432	OUP
2006	408	OUP
2007	449	OUP
2008	409	OUP
2009	445	OUP
2010	349	OUP
Total	7433	
Elsevier	3440	
OUP	3993	
Periode 1995-2002	4339	(dont 3440 elsevier et 899 OUP)
Periode 2003-2010	3094	(dont 0 elsevier et 3094 OUP)



# Prétraitement des textes (WPO)

```
<TEI>
<teiHeader>...</teiHeader>
<body>
<text>
<div type="div1">
<head subtype="level1"><w xml:id="d1e230">Introduction</w></head>
<p><w xml:id="d1e234">L'</w><w xml:id="d1e237">usage</w> <w xml:id="d1e240">des</w> <w xml:id="d1e243">«</w> <w xml:id="d1e247">connecteurs</w> <w xml:id="d1e250">»</w> <w xml:id="d1e254">ou</w> <w xml:id="d1e257">«</w> <w xml:id="d1e261">mots</w>
<w xml:id="d1e264">de</w> <w xml:id="d1e268">discours</w> <w xml:id="d1e271">»</w> <pc xml:id="d1e274">(</pc><w xml:id="d1e277">Ducrot</w> <w xml:id="d1e281">1980</w><pc xml:id="d1e283">)</pc> <w xml:id="d1e286">représente</w> <w xml:id="d1e289">un
</w> <w xml:id="d1e292">des</w> <w xml:id="d1e296">moyens</w> <w xml:id="d1e300">linguistiques</w> <w xml:id="d1e304">dont</
w> <w xml:id="d1e307">dispose</w> <w xml:id="d1e310">tout</w> <w xml:id="d1e314">locuteur</w> ... </p>
</text>
</body>
<stdf>
<spanGrp type="wordForms">
<span target="#d1e234" pos="DET:ART" lemma="le" corresp="dictionnaire_TLFiCategoriesTEI#l19339"/>
<span target="#d1e237" pos="NOM" lemma="usage" corresp="dictionnaire_TLFiCategoriesTEI#l60800"/>
<span target="#d1e240" pos="PRP:det" lemma="du" corresp="dictionnaire_TLFiCategoriesTEI#l78608"/>
<span target="#d1e243" pos="PUN:cit" lemma="«" corresp=""/>
<span target="#d1e247" pos="NOM" lemma="connecteur" corresp="DicoWiktionaryCategoriesTEI#l10412"/>
</spanGrp>
<spanGrp type="candidatsTermes">
<span target="#d1e230" lemma="introduction" ana="#DM2 #DAoff" corresp="#entry-479294"/>
<span target="#d1e237" lemma="usage" ana="#DM3 #DAoff" corresp="#entry-220576"/>
<span target="#d1e247" lemma="connecteur" ana="#DM4 #DAon" corresp="#entry-386124"/>
<span target="#d1e261 #d1e264 #d1e268" lemma="mot de discours" ana="#DM3 #DAoff" corresp="#entry-1039094"/>
<span target="#d1e268" lemma="discours" ana="#DM4 #DAon" corresp="#entry-454808"/>
<span target="#d1e277" lemma="ducrot" ana="#DM3 #DAoff" corresp="#entry-395608"/>
<span target="#d1e296 #d1e300" lemma="moyen linguistique" ana="#DM3 #DAoff" corresp="#entry-1613170"/>
<span target="#d1e300" lemma="linguistique" ana="#DM1 #DAoff" corresp="#entry-47530"/>
<span target="#d1e314" lemma="locuteur" ana="#DM4 #DAon" corresp="#entry-490000"/>
</spanGrp>
</stdf>
</TEI>
```

Référence des tokens

Annotation en traits sémantiques  
extraits du TLFi et du Wiktionary

Texte

POS et semantic tagging

Annotations

Occurrences de  
termes et résultats  
de désambiguïsation

Références croisées

Désambiguïsation manuelle  
(DM) et automatique (DA)

Correspondance avec la terminologie  
extraite par TermSuite

# Prétraitement des textes (WPO)

- affichage des candidats (puces vertes)
- des expressions figées et semi-figées de la langue du lexique transdisciplinaire (triangles oranges) ou non (carrés bleus).

## 2.1 Faits • [marquants] en • [évaluation] de la TA

L'histoire de l' • [évaluation] en TA remonte à ses débuts. On dit parfois, en plaisantant ▲ [à peine], qu'elle • [a • [donné lieu] à] plus de] publications que la TA elle • [-même]. On a toujours distingué entre • [ • [évaluations] • [externes]], jugeant la • [qualité des résultats] sur des • [ • [critères] • [linguistiques]] ( • [grammaticalité], fidélité, etc.) ou sur des • [critères] d' • [usage] (productivité, coût), et • [évaluations] internes, jugeant la conception des systèmes (architecture • [linguistique] et architecture calculatoire) et leurs perspectives d'amélioration et d'extension à de • [nouvelles] • [langues], à de nouveaux • [types] de documents, et à de • [nouvelles] tâches ( • [e.g. de l'assimilation à la dissémination]).

Fin 1966, le • [rapport • [ALPAC]] ( • [ALPAC], 1966), fondé sur une • [évaluation] contestable et contestée des systèmes de TA d'alors -sauf curieusement la version la plus • [récente] du système le plus proche (le Georgetown • [Automatic] • [Translation] ou GAT) - eut une conséquence • [importante], l'arrêt presque total du financement de la recherche en TA aux USA pour près de 20 ans (jusqu'à 1985), et ▲ [par • [ricochet]] en • [Angleterre] et au • [Japon], ce • [dernier] pays y revenant vers 1980, à cause des besoins très • [importants] en automatisation de la • [traduction] liés à l'isolement du japonais, à sa difficulté, et à son importance • [commerciale] • [grandissante]. La TA • [opérationnelle] continua cependant, et les systèmes russe-anglais GAT puis • [Systran] furent régulièrement évalués (à la Wright-Patterson Air Force Base et à Ispra, EURATOM) selon deux • [critères], la • [qualité • [linguistique]] et l'utilité pratique, avec des résultats totalement opposés : lors du séminaire organisé en 1972 à • [Austin], • [Texas], on • [parlait de] 2/20 (E) en • [qualité • [linguistique]] et de 18/20 (A) en utilité. ▲ [ • [En tout cas]], le • [rapport • [ALPAC]] jeta les bases des protocoles d' • [ • [évaluation] • [subjective]] (reposant sur des • [jugements • [humains]] et non • [sur des mesures] de performance en temps, ni sur des • [comparaisons] • [automatiques] avec des • [références]).



# Prétraitement des textes (WPO)

- affichage des candidats (puces vertes)
- des expressions figées et semi-figées de la langue du lexique transdisciplinaire (triangles oranges) ou non (carrés bleus).

## 2.1 Faits • [marquants] en • [évaluation] de la TA

L'histoire de l' • [évaluation] en TA remonte à ses débuts. On dit parfois, en plaisantant ▲ [à peine], qu'elle • [a • [donné lieu] à] plus de] publications que la TA elle • [-même]. On a toujours distingué entre • [ • [évaluations] • [externes]], jugeant la • [qualité des résultats] sur des • [ • [critères] • [linguistiques]] ( • [grammaticalité], fidélité, etc.) ou sur des • [critères] d' • [usage] (productivité, coût), et • [évaluations] internes, jugeant la conception des systèmes (architecture • [linguistique] et architecture calculatoire) et leurs perspectives d'amélioration et d'extension à de • [nouvelles] • [langues], à de nouveaux • [types] de documents, et à de • [nouvelles] tâches ( • [e.g]. de l'assimilation à la dissémination).

Fin 1966, le • [rapport • [ALPAC]] ( • [ALPAC], 1966), fondé sur une • [évaluation] contestable et contestée des systèmes de TA d'alors -sauf curieusement la version la plus • [récente] du système le plus proche (le Georgetown • [Automatic] • [Translation] ou GAT) - eut une conséquence • [importante], l'arrêt presque total du financement de la recherche en TA aux USA pour près de 20 ans (jusqu'à 1985), et ▲ [par • [ricochet]] en • [Angleterre] et au • [Japon], ce • [dernier] pays y revenant vers 1980, à cause des besoins très • [importants] en automatisation de la • [traduction] liés à l'isolement du japonais, à sa difficulté, et à son importance • [commerciale] • [grandissante]. La TA • [opérationnelle] continua cependant, et les systèmes russe-anglais GAT puis • [Systran] furent régulièrement évalués (à la Wright-Patterson Air Force Base et à Ispra, EURATOM) selon deux • [critères], la • [qualité • [linguistique]] et l'utilité pratique, avec des résultats totalement opposés : lors du séminaire organisé en 1972 à • [Austin], • [Texas], on • [parlait de] 2/20 (E) en • [qualité • [linguistique]] et de 18/20 (A) en utilité. ▲ [ • [En tout cas]], le • [rapport • [ALPAC]] jeta les bases des protocoles d' • [ • [évaluation] • [subjective]] (reposant sur des • [jugements • [humains]] et non • [sur des mesures] de performance en temps, ni sur des • [comparaisons] • [automatiques] avec des • [références]).

# Extraction de connaissances (WP2)

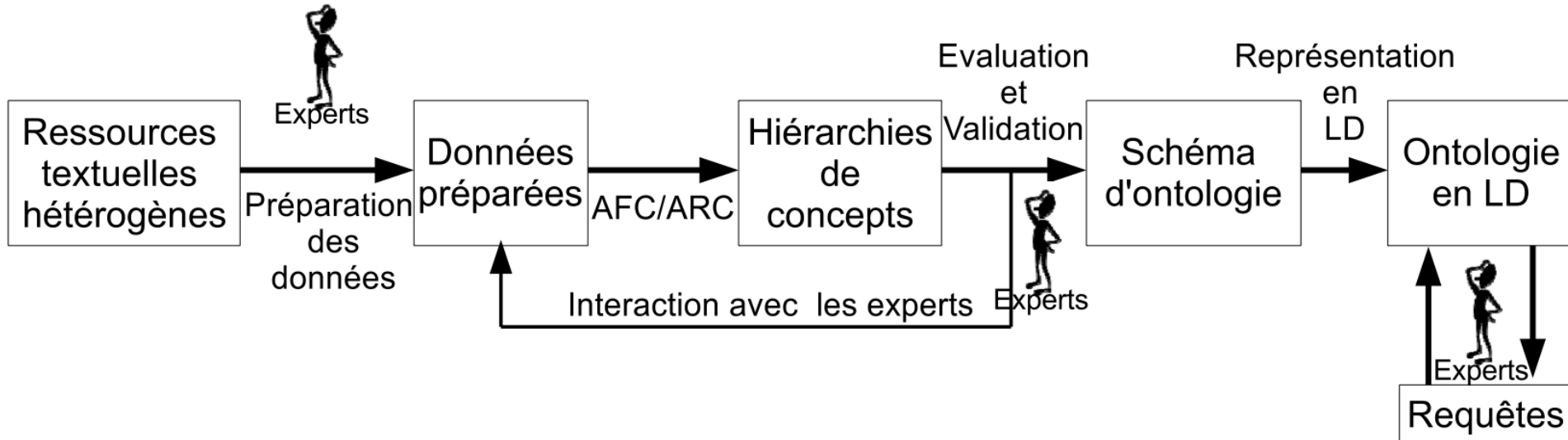
- Mettre en œuvre des méthodes de fouille de données symboliques applicables aux textes
  - Fouille de motifs, de séquences, d'arbres ou de graphes
  - Classification par Analyse Formelle de Concepts

# Extraction de connaissances (WP2)

- Les méthodes
  - Extraction de motifs
    - Réduction du nombre de motifs
    - Validation de l'occurrence du terme « argument »
      - Motif positif : [*sdrt, être, argument*]
      - Motif négatif : [*trancher, pas, ne, argument, permettre, décisif, position, avoir*]
    - Extraction de motifs séquentiels pour l'identification de relations
  - Analyse formelle de concept pour la conceptualisation du domaine

# Extraction de connaissances (WP2)

- Mise en place d'une plateforme développée sous GATE
- Articulation de différents niveaux de traitement : TAL, FdD, RC



# Prétraitement sous Gate (Alzheimer)

Language Resources

Processing Resources

Simple PubMed Extractor

Datastores

- Getting ids...  
65466 abstracts found  
done in 3.437s

- Getting xml files from ids...  
<http://www.ncbi.nlm.nih.gov/pubmed/25702360?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25702360.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25693568?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25693568.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25665290?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25665290.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25643578?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25643578.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25597360?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25597360.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25590418?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25590418.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25575135?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25575135.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25575131?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25575131.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25574508?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25574508.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25558551?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25558551.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25552162?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25552162.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25550561?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25550561.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25546749?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25546749.xml  
<http://www.ncbi.nlm.nih.gov/pubmed/25543114?dopt=XML>  
writing xml file: /home/lmelo/Bureau/AlzheimerDisease/25543114.xml

# Annotation par des ressources terminologiques

er 8.0 build 4825  
ols Help

Messages MetaMap 25702360.txt\_00...

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

Long-term efficacy and toxicity of cholinesterase inhibitors in the treatment of Alzheimer disease. Though the symptoms of Alzheimer disease go on for years, the phase 3 trials of the cholinesterase inhibitors (ChEIs), the current mainstay of symptomatic pharmacotherapy for this condition, were typically of only 3- to 6-months' duration. We have limited data on long-term (that is, a year or more) therapy with these agents. In this review, we explore the available information on the biological and clinical effects of long-term ChEI therapy, what happens when these agents are discontinued, and examine what others have recommended. An individualized approach to deciding on whether to carry on with a ChEI should be taken. If continued, treatment goals should be clarified and patients monitored over time, for both drug-related benefits and adverse effects.

Type	Set	Start	End	Id	
Temporal Concept		0	9	1149	{ ConceptId=C0443252, PreferredName=Long-term, String=Long-term, id=1149}
Pharmacologic Substance		35	60	1151	{ ConceptId=C0008425, PreferredName=Cholinesterase Inhibitors, String=cholinesterase inhibit
Disease or Syndrome		81	98	1142	{ ConceptId=C0002395, PreferredName=Alzheimer's Disease, String=Alzheimer disease, id=1142}
Disease or Syndrome		123	140	1140	{ ConceptId=C0002395, PreferredName=Alzheimer's Disease, String=Alzheimer disease, id=1140}
Temporal Concept		151	156	1147	{ ConceptId=C0439234, PreferredName=year, String=years, id=1147}
Temporal Concept		162	169	1148	{ ConceptId=C0439561, PreferredName=Phase 3, String=phase 3, id=1148}
Pharmacologic Substance		184	209	1146	{ ConceptId=C0008425, PreferredName=Cholinesterase Inhibitors, String=cholinesterase inhibit
Temporal Concept		223	230	1166	{ ConceptId=C0521116, PreferredName=Current (present time), String=current, id=1166}
Therapeutic or Preventive Procedure		255	270	1164	{ ConceptId=C0013216, PreferredName=Pharmacotherapy, String=pharmacotherapy, id=1164}
Temporal Concept		322	328	1168	{ ConceptId=C0439231, PreferredName=month, String=months, id=1168}
Temporal Concept		330	338	1167	{ ConceptId=C0449238, PreferredName=Duration (temporal concept), String=duration, id=1167}
Temporal Concept		364	373	1156	{ ConceptId=C0443252, PreferredName=Long-term, String=long-term, id=1156}
Temporal Concept		386	390	1155	{ ConceptId=C0439234, PreferredName=year, String=year, id=1155}
Chemical Viewed Functionally		419	425	1161	{ ConceptId=C0450442, PreferredName=Agent, String=agents, id=1161}
Temporal Concept		522	531	1122	{ ConceptId=C0443252, PreferredName=Long-term, String=long-term, id=1122}
Chemical Viewed Functionally		570	576	1125	{ ConceptId=C0450442, PreferredName=Agent, String=agents, id=1125}
Patient or Disabled Group		781	789	1138	{ ConceptId=C0030705, PreferredName=Patients, String=patients, id=1138}
Temporal Concept		805	809	1132	{ ConceptId=C0040223, PreferredName=Time, String=time, id=1132}
Pharmacologic Substance		820	824	1135	{ ConceptId=C0013227, PreferredName=Pharmaceutical Preparations, String=drug, id=1135}

19 Annotations (0 selected) Select:

Document Editor Initialisation Parameters Relation Viewer

- Activity
- Biomedical\_Occupation\_or\_
- Chemical\_Viewed\_Function
- Conceptual\_Entity
- Disease\_or\_Syndrome
- Finding
- Functional\_Concept
- Gene\_or\_Genome
- Health\_Care\_Activity
- Idea\_or\_Concept
- Intellectual\_Product
- MetaMap
- Organic\_Chemical
- Patient\_or\_Disabled\_Group
- Pharmacologic\_Substance
- Qualitative\_Concept
- Quantitative\_Concept
- Research\_Activity
- Sentence
- SpaceToken
- Spatial\_Concept
- Split
- Temporal\_Concept
- Therapeutic\_or\_Preventive
- Token
- Original\_markup



# Annotation syntaxique

The screenshot displays the GATE (General Architecture for Text Engineering) software interface. The main window shows a text document with syntactic annotations. A 'Syntax tree viewer' window is open, displaying a parse tree for the sentence: "Long-term efficacy and toxicity of cholinesterase inhibitors in the treatment of Alzheimer disease".

The parse tree structure is as follows:

- ROOT
  - NP
    - PP (Long-term)
    - PP (efficacy and toxicity of)
    - NP (cholinesterase inhibitors in the treatment of Alzheimer disease)
      - NP (cholinesterase inhibitors)
      - PP (in the treatment of)
      - NP (Alzheimer disease)

The 'Syntax tree viewer' window also shows the following text: "Long-term efficacy and toxicity of cholinesterase inhibitors in the treatment of Alzheimer disease".

The GATE interface includes a menu bar (File, Options, Tools, Help), a toolbar, and a sidebar with a tree view of applications (SyntaxAnalysis, MetaMap, Preprocessing, CollectionOfCorpus) and language resources (GATE Corpus\_000F4, 25702360.txt\_00071, etc.). The main window has tabs for Messages, SyntaxAnalysis, and 25702360.txt\_00... The SyntaxAnalysis tab is active, showing a list of annotation sets and a list of annotations. The text in the main window is: "Long-term efficacy and toxicity of cholinesterase inhibitors in the treatment of Alzheimer disease. Though the symptoms of Alzheimer disease go on for years, the phase 3 trials of the cholinesterase inhibitors (ChEIs), the current mainstay of symptomatic pharmacotherapy for this condition, were typically of only 3- to 6-months' duration. We have limited data on long-term (that is, a year or more) therapy with these agents. In this review, we explore the available information on the biological and clinical effects of long-term ChEI therapy, what happens when these agents are discontinued, and examine what others have recommended an individualized approach to deciding on whether to carry on with a ChEI should be taken. If continued, treatment goals should be clarified and patients monitored over time, for both drug-related benefits and adverse effects."

# Extraction de relations

Developer 8.0 build 4825  
is Tools Help

Messages 25496901.txt.xml...

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

Wilson's disease and other neurological copper disorders.  
The copper metabolism disorder Wilson's disease was first defined in 1912. **Wilson's disease can present with hepatic and neurological deficits, including dystonia and parkinsonism.** Early-onset presentations in infancy and late-onset manifestations in adults older than 70 years of age are now well recognised. Direct genetic testing for ATP7B mutations are increasingly available to confirm the clinical diagnosis of Wilson's disease, and results from biochemical and genetic prevalence studies suggest that Wilson's disease might be much more common than previously estimated. Early diagnosis of Wilson's disease is crucial to ensure that patients can be started on adequate treatment, but uncertainty remains about the best possible choice of medication. Furthermore, Wilson's disease needs to be differentiated from other conditions that also present clinically with hepatolenticular degeneration or share biochemical abnormalities with Wilson's disease, such as reduced serum ceruloplasmin concentrations. Disordered copper metabolism is also associated with other neurological conditions, including a subtype of axonal neuropathy due to ATP7A mutations and the late-onset neurodegenerative disorders Alzheimer's disease and Parkinson's disease.

Features	
tion, actorType=Disease_or_Syndrome, goalId=1598, goalString=Parkinsonian Disorders, goalType=Disease_or_Syndrome, verb=present}	
tion, actorType=Disease_or_Syndrome, goalId=1619, goalString=Neurologic Deficits, goalType=Finding, verb=present}	
tion, actorType=Disease_or_Syndrome, goalId=1597, goalString=Dystonia, goalType=Sign_or_Symptom, verb=present}	

- Laboratory\_or\_Test\_R
- Mental\_Process
- Mental\_or\_Behavioral
- MetaMap
- Organism\_Attribute
- Organism\_Function
- Patient\_or\_Disabled\_G
- Pharmacologic\_Substa
- Qualitative\_Concept
- Quantitative\_Concept
- Sentence
- Sign\_or\_Symptom
- Social\_Behavior
- SpaceToken
- Split
- SyntaxTreeNode
- Temporal\_Concept
- Token
- dsyn\_dsyn\_
- dsyn\_fndg\_
- dsyn\_sosy\_
- Original markups

3 Annotations (0 selected) Select:

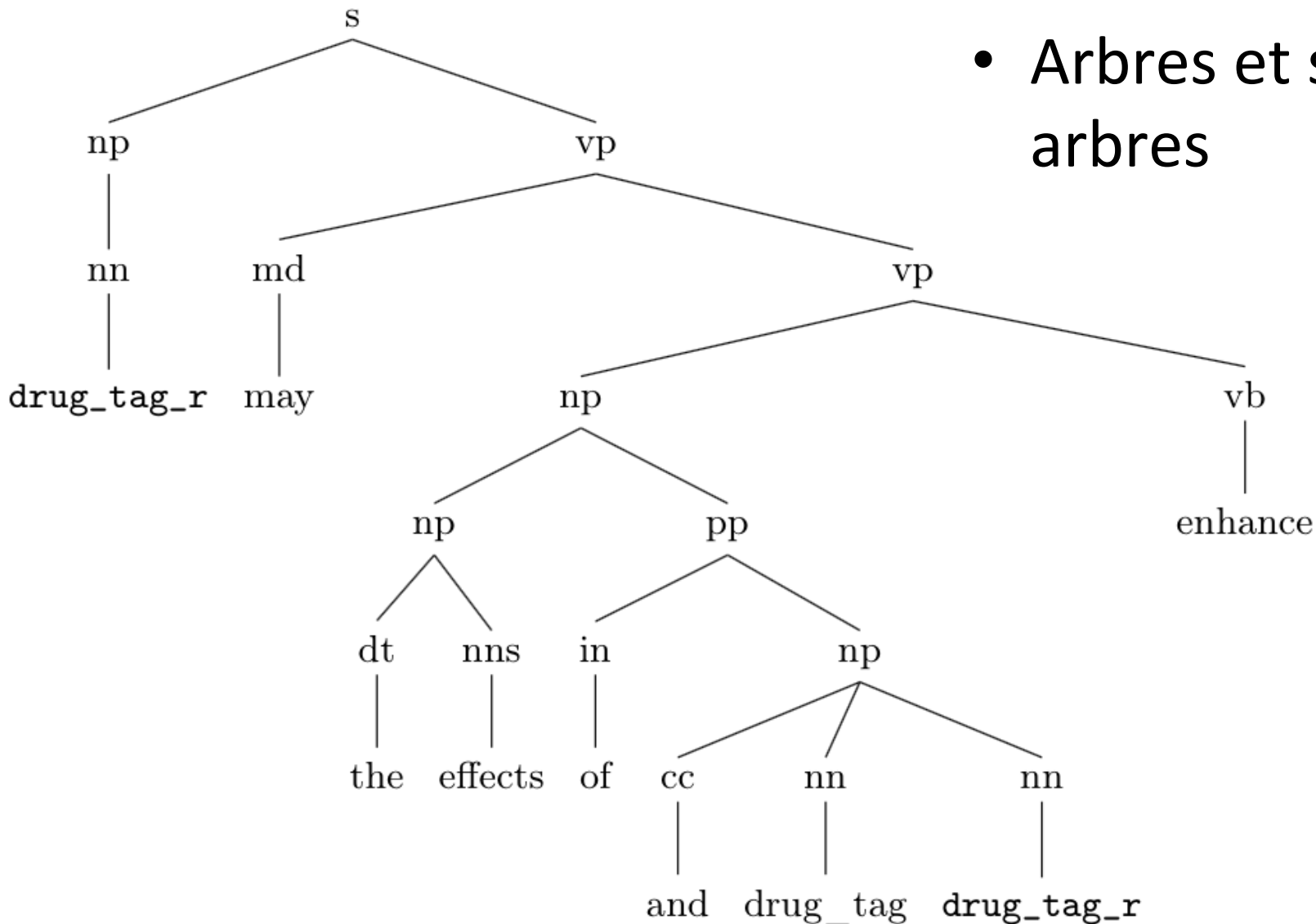
Document Editor Initialisation Parameters Relation Viewer

# Extraction de relations

The screenshot shows a software window titled "d 4825" with a toolbar containing icons for a red cube, a green cross, and a green checkmark. Below the toolbar is a "Messages" section with three document icons labeled "25496901.txt.xm...", "25471565.txt.xm...", and "25693568.txt.xm...". The main interface has a menu bar with "Annotation Sets", "Annotations List", "Annotations Stack", "Co-reference Editor", and "Text". The main text area contains a paragraph of text with a highlighted sentence: "Alzheimer's disease (AD) is a severe age-related neurodegenerative disorder characterized by accumulation of amyloid- $\beta$  plaques and neurofibrillary tangles, synaptic and neuronal loss, and cognitive decline." Below the text is a "Features" section with a single line of text: "=Alzheimer's Disease, actorType=Disease\_or\_Syndrome, goalId=1900, goalString=Neurodegenerative Disorders, goalType=Disease\_or\_Syndrome, relationOrigin=Copula". At the bottom, there is a status bar showing "1 Annotations (0 selected) Select:" and a menu bar with "Document Editor", "Initialisation Parameters", and "Relation Viewer". On the right side, there is a vertical list of colored checkboxes with labels: Man, Med, Men, Men, Met, Mol, Nat, Nucl, Org., Org., Org., Qua, Qua, Sen, Spa, Spa, Split, Synt, Tem, Ther, Toki, and dsyr.

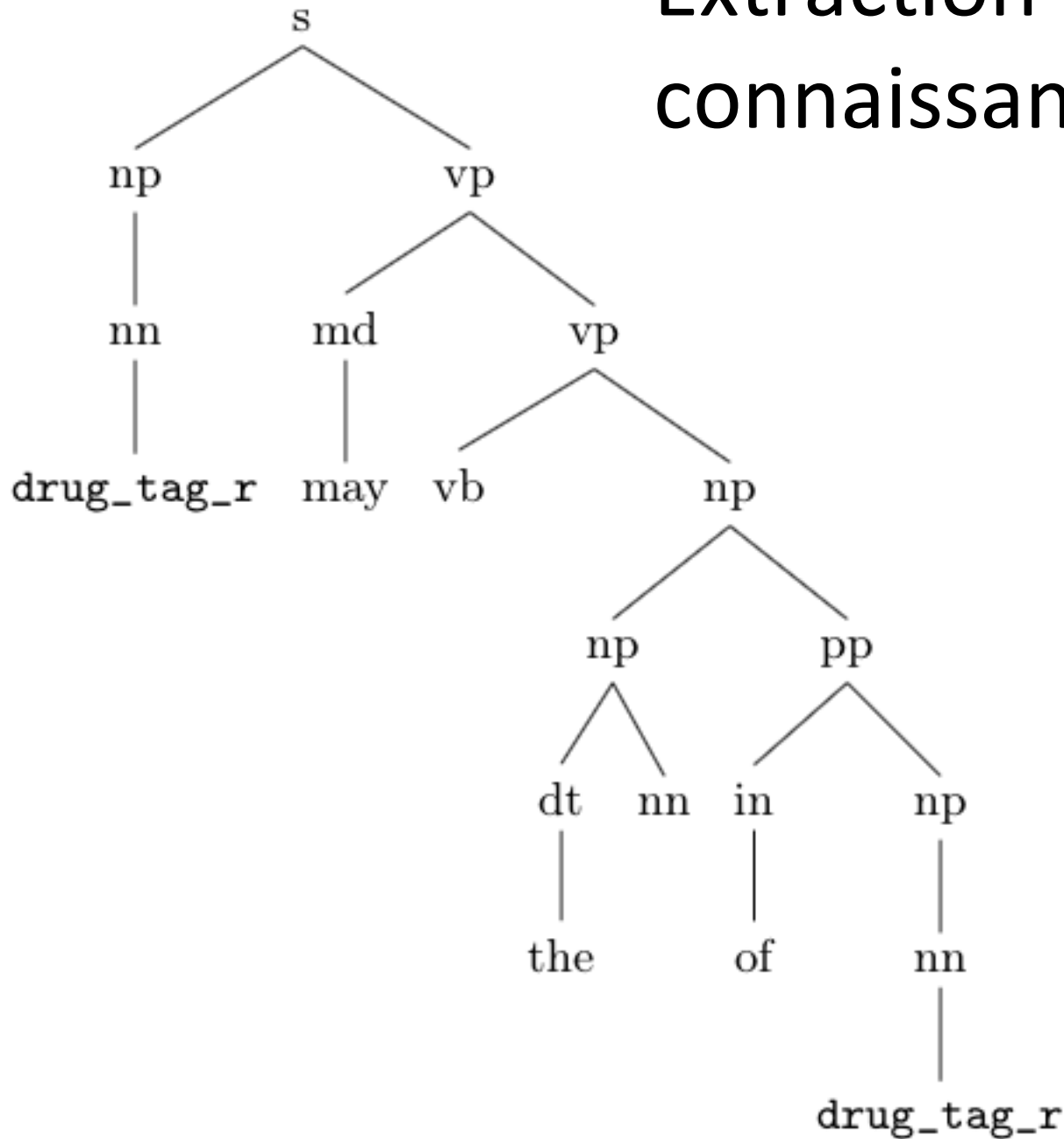
# Extraction de connaissances (WP2)

- Arbres et sous-arbres



# Extraction de connaissances (WP2)

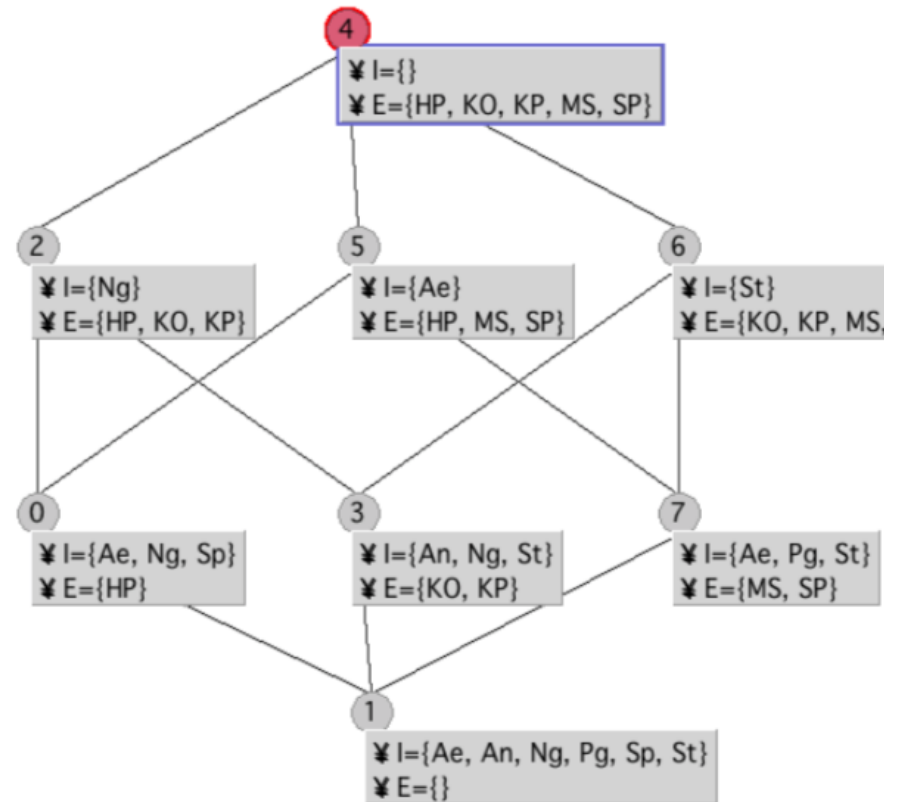
Arbres et sous-arbres



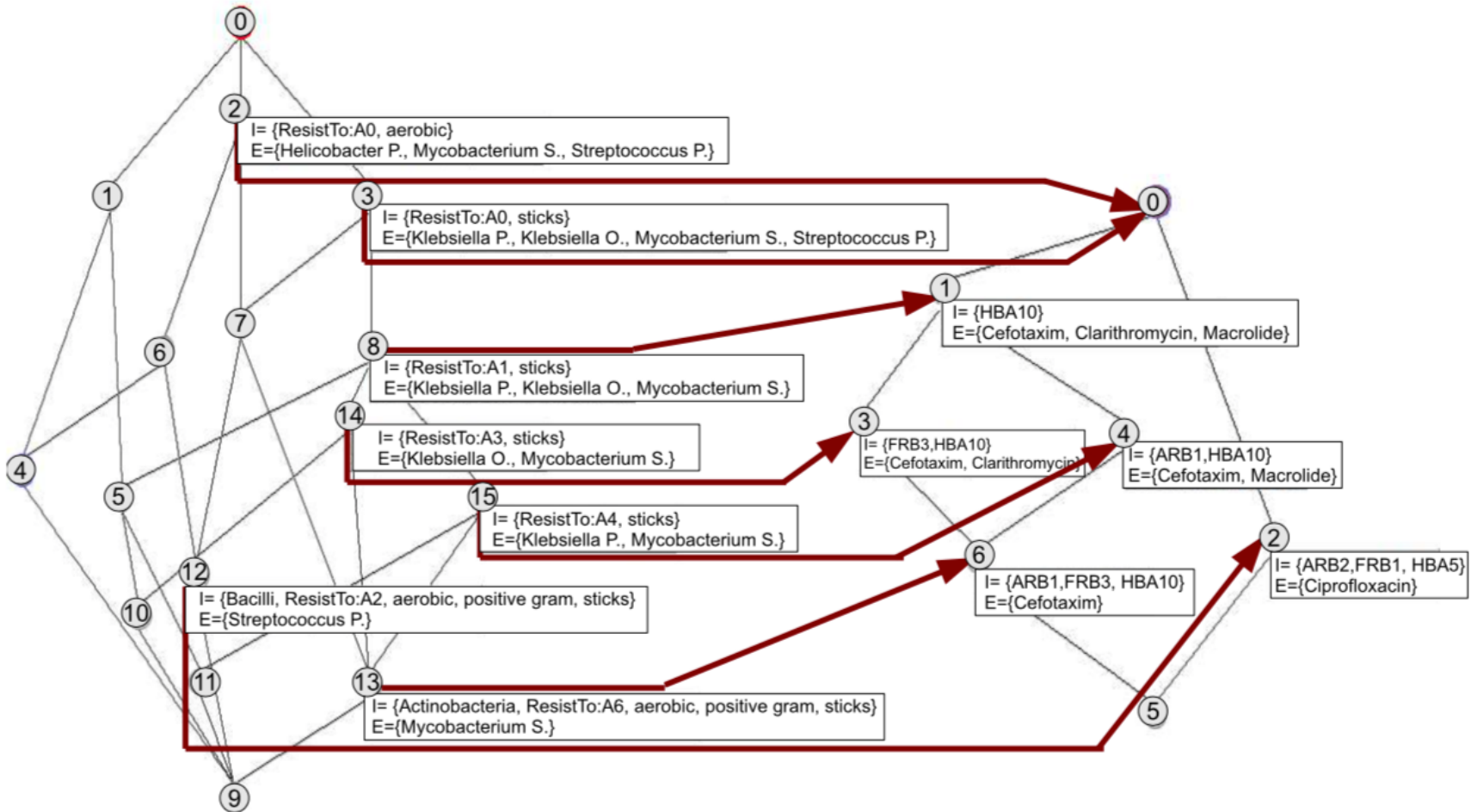


# Analyse formelle de concepts

Bacteria	spherical (Sp)	sticks (St)	negative gram (Ng)	positive gram (Pg)	aerobic (Ae)	anaerobic (An)
Helicobacter P. (HP)	x		x		x	
Klebsiella P. (KP)		x	x			x
Mycobacterium S. (MS)		x		x	x	
Streptococcus P. (SP)		x		x	x	
Klebsiella O. (KO)		x	x			x



# Objets complexes en AFC





# Analyse diachronique et classification automatique des textes



# CONTRAINTES ÉMERGENTES EN ANALYSE DES DONNÉES TEXTUELLES

- L'analyse des données textuelles nécessite de s'adapter aux nouvelles formes d'information disponibles en ligne.
- Cela implique de prendre en compte des techniques qui supportent :
  - Les données volumineuses, éparses et fortement multidimensionnelles,
  - Le traitement des données rares, similaires et/ou déséquilibrées,
  - Le traitement des données changeantes,
  - Les interactions multiples entre les sources,
- Une des tâches principales de l'analyse est la classification/discrimination des données.
- Les méthodes/distances classiques s'adaptent mal à ces contraintes.

# SOLUTIONS APPORTÉES

- Réviser la notion de distance classique en examinant de meilleurs compromis entre la généralité et la discrimination :
    - Les distances euclidiennes sont inadaptées aux données éparses et/ou multidimensionnelles,
    - Les distances statistiques comme le Khi 2 privilégient la discrimination par rapport à la généralité.
  - Chercher des solutions alternatives en s'inspirant du domaine de la recherche d'information :
    - Rappel (ou discrimination),
    - Précision (ou généralité),
    - Mesure F (ou compromis discrimination/généralité).
  - Conserver la possibilité de mener des analyses non supervisées.
- => **Théorie de la maximisation de l'étiquetage.**

# CLUSTERING ET MAXIMISATION D'ÉTIQUETAGE

- Clustering
  - Permet d'organiser l'information en thématiques si celles-ci ne sont pas présentes dans le corpus (ou ne sont pas construites selon les mêmes normes)
  - Permet de simplifier la visualisation des résultats de recherche
  - Permet l'analyse des changements de sujets dans le cas d'une approche incrémentale
- Méthodes employées
  - Méthodes neuronales statiques
  - Optimisation des modèles et méthodes neuronales incrémentales (IGNG-F) basées sur la métrique de maximisation d'étiquetage

# CLUSTERING ET MAXIMISATION D'ÉTIQUETAGE

Soit un ensemble de groupes  $C$  issu d'une méthode de regroupement sur un ensemble de données  $D$  représentées par un ensemble de traits descriptifs  $F$ , la maximisation des traits est une métrique qui favorise les groupes qui maximisent la F-mesure de traits, moyenne harmonique entre :

**Rappel de trait**

$$FR \downarrow c(f) = \frac{\sum_{d \in c} W \downarrow d \uparrow f}{\sum_{c' \in C} \sum_{d \in c'} W \downarrow d \uparrow f} \equiv P(c|f)$$

**Prépondérance de trait**

$$FP \downarrow c(f) = \frac{\sum_{d \in c} W \downarrow d \uparrow f}{\sum_{f' \in F} \sum_{d \in c} W \downarrow d \uparrow f'} \equiv P(f|c)$$

Le principe de maximisation des traits permet de sélectionner par exemple les traits (i.e. termes) caractéristiques dans un graphe

# ADAPTATION DE LA MÉTRIQUE DE MAXIMISATION D'ÉTIQUETAGE POUR LA SÉLECTION DE VARIABLES

Le processus de maximisation d'étiquetage peut être appliqué sur les classes aussi bien que sur des clusters dès lors qu'il est seulement dépendant des données associées. C'est un processus sans paramètres.

L'ensemble  $S_c$  des variables qui sont caractéristiques d'une classe donnée  $c$  appartenant à un ensemble de classe  $C$  se traduit par :

$$S_c = \{f \in F \mid FF_c(f) > FF(f) \text{ and } FF_c(f) > FF_D\}$$

où  $FF_c(f) = \sum_{c' \in C \mid c' \neq c} FF_{c'}(f) / |C|$  et  $FF_D = \sum_{f \in F} FF(f) / |F|$ .

Et  $C_f$  représente un sous-ensemble de  $C$  aux classes dans lesquelles la variable  $f$  est représentée.

Enfin, l'ensemble de toutes les variables sélectionnées  $S_C$  est le sous-ensemble de  $F$  défini comme :

$$S_C = \bigcup_{c \in C} S_c$$

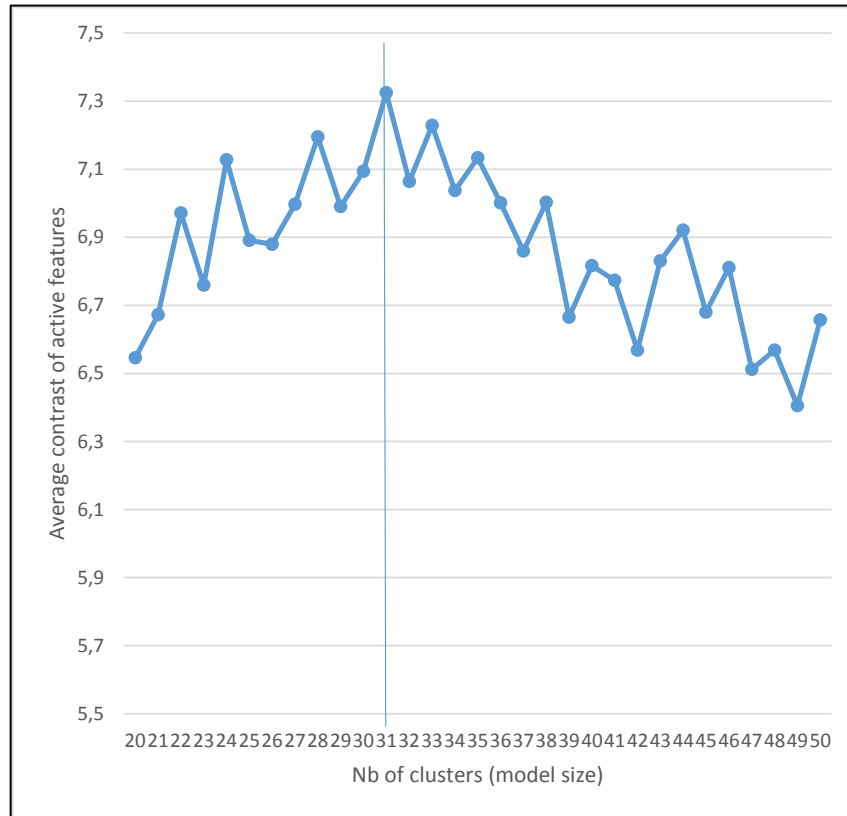
# MÉTRIQUE F-MAX BASÉE SUR LE CONTRASTE

Le contraste ou gain d'information consiste à trouver la force de la relation entre une variable et une classe. Pour une variable  $f$  associée à une classe  $C$ , il s'exprime comme suit:

$$C \downarrow c (f) = (FF \downarrow c (f) / FF (f)) \uparrow k$$

- Un contraste  $> 1$  met en évidence un comportement actif de la variable dans la classe,
- Un contraste  $< 1$  est lié à un comportement passif de la variable dans la classe,
- Le facteur de magnification  $k$  peut être utilisé en classification pour mieux séparer les classes, le cas échéant (séparation non linéaire).

# CLUSTERING ET MAXIMISATION D'ÉTIQUETAGE



Exemple de traitement non supervisé à partir d'un ensemble de notices extraites du corpus ISTEEX sur un sujet général (ici optoélectronique) :

**Phase 1** : analyse des textes

**Phase 2** : clustering avec choix de la méthode optimale et extraction du modèle optimal

**Phase 3** : visualisations

Le meilleur modèle est isolé en exploitant la variation d'index basés sur le contraste moyen des traits actifs des clusters (ici un modèle à 31 clusters)



# CLUSTERING ET MAXIMISATION D'ÉTIQUETAGE

## Cluster 6

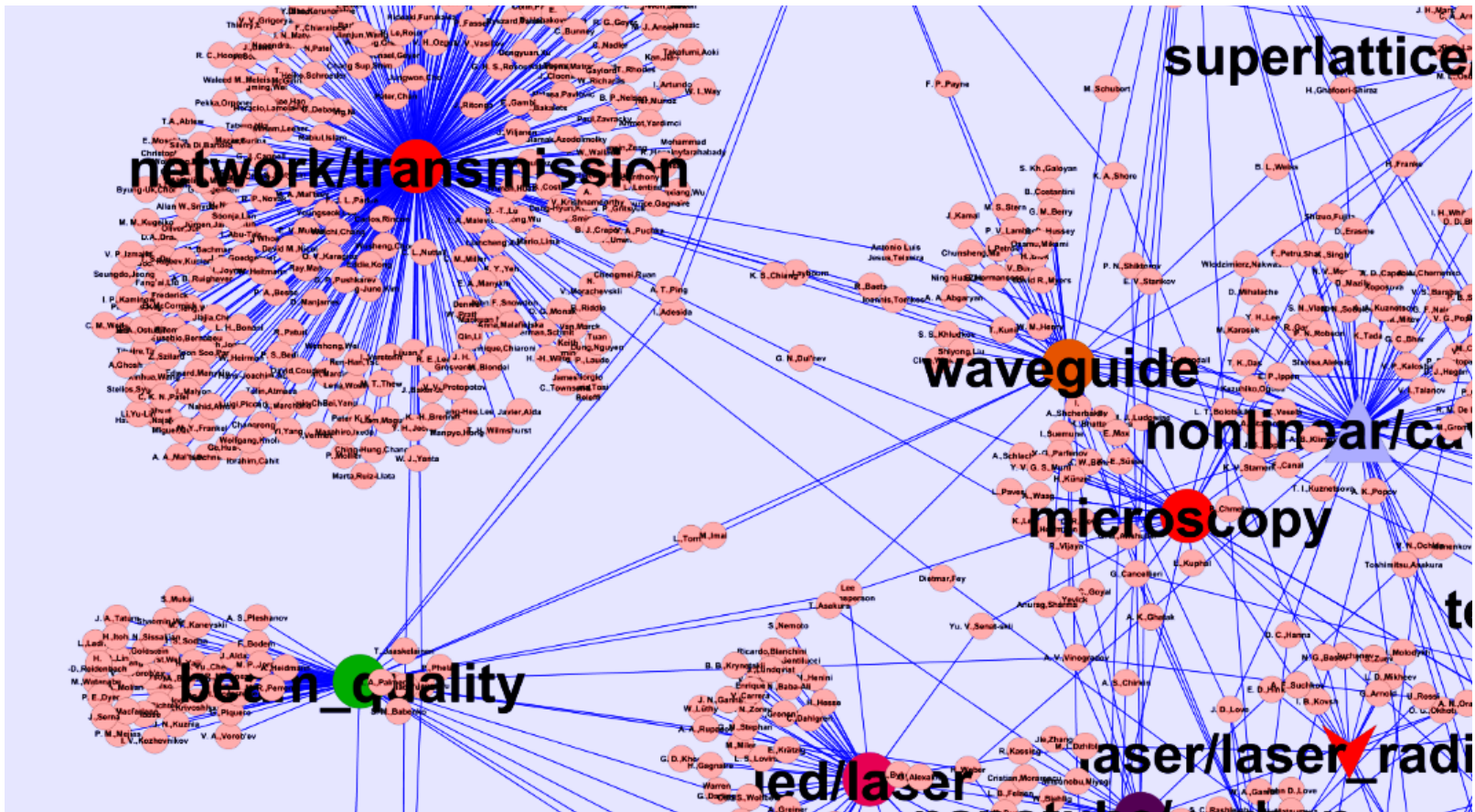
```
13.320440 temperature
8.408364 temperature_dependence
7.580113 room_temperature
6.336815 dependence
5.835384 band
5.031722 temperature_range
4.623655 room
4.283682 emission
4.227297 range
4.170750 activation_energy
4.098943 peak
3.803711 edge
```

## Cluster 14

```
14.105946 crystal
4.459125 oxide
3.885909 crystal_growth
3.226209 alxga
3.136092 iron
3.062750 transition
2.815242 defect
2.807345 capacitance
2.741107 transport_properties
2.702896 hall_measurements
2.603363 inp
2.584676 crystal_quality
2.549993 photonic_crystals
```

Le profil thématique d'un cluster (sujet(s) couvert(s)) est mis automatiquement en évidence aux utilisateurs à partir de ses traits actifs de plus fort contraste

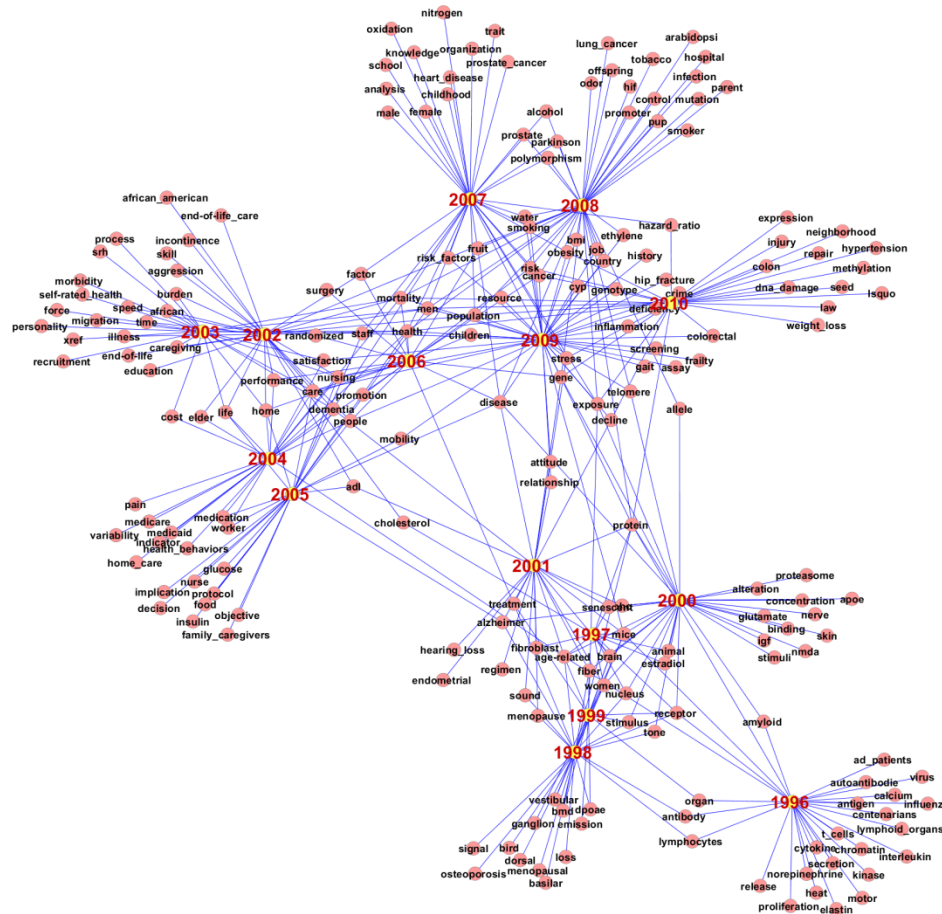
# CLUSTERING ET MAXIMISATION D'ÉTIQUETAGE



Les interactions auteurs-sujets et la centralité des auteurs peuvent être directement caractérisés sur un graphe de contraste construit à partir des résultats du clustering

# GRAPHE MOTS-ANNÉES ET MAXIMISATION D'ÉTIQUETAGE

Corpus ISTE  
« Vieillesse »

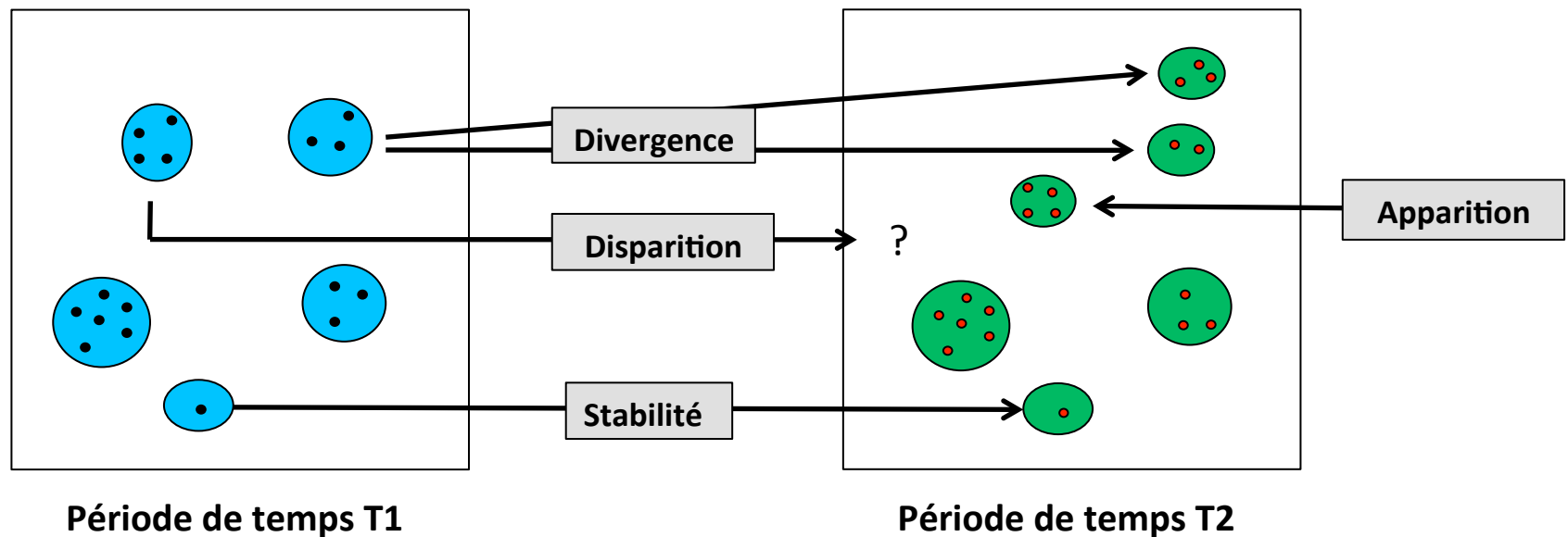


On peut voir dans les années 2003-2006 l'apparition de termes comme "nurse, nursing, home\_care, medicare, family caregivers, home, satisfaction, home care..." qui marquent le développement du maintien des personnes âgées à domicile

# ANALYSE DIACHRONIQUE DE LA RECHERCHE

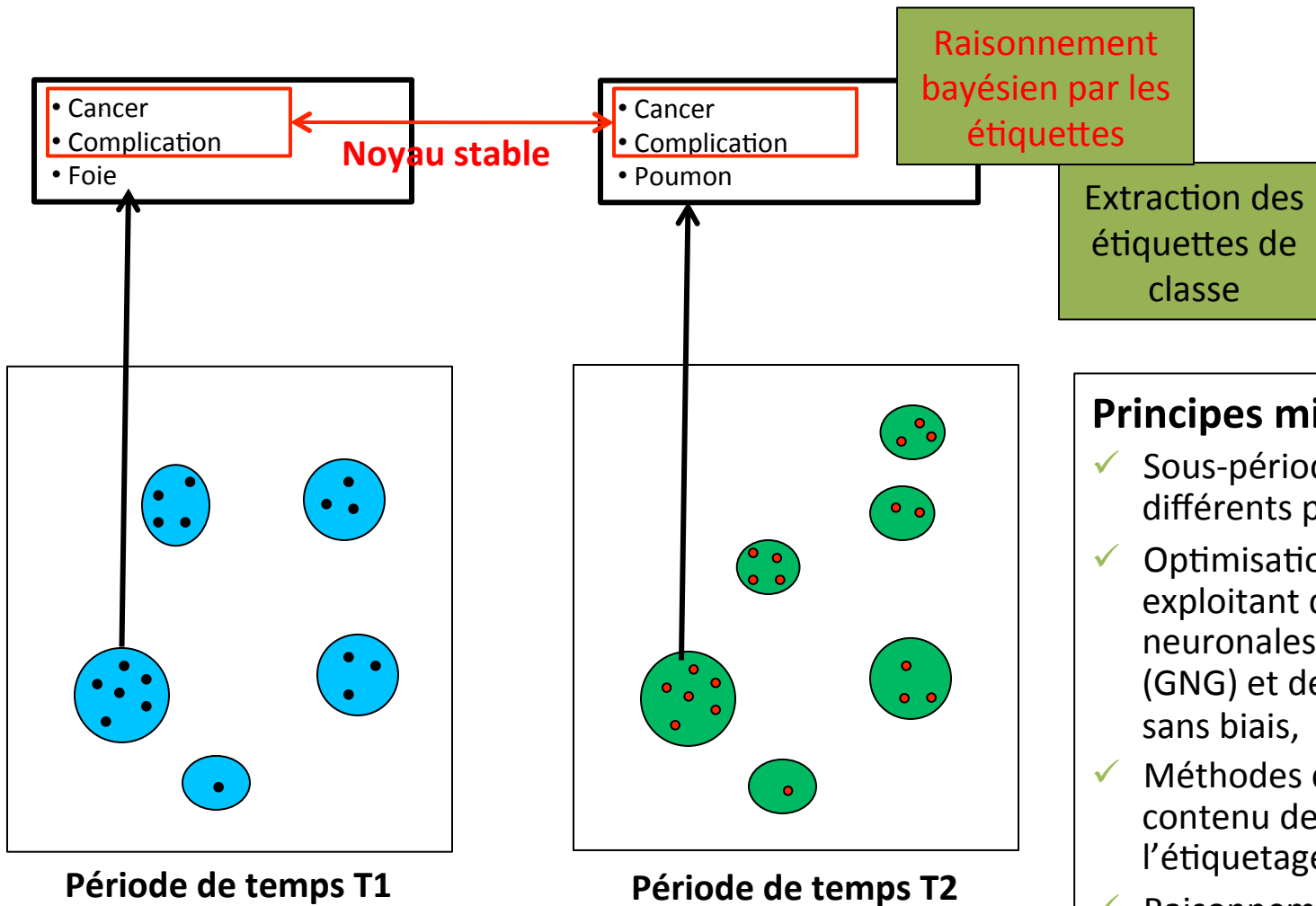
## Buts :

- ✓ Automatiser le processus d'analyse par pas de temps (analyse diachronique) de l'évolution des thèmes de recherches en exploitant les capacités du paradigme MVDA et celles de la maximisation.



Expérience basée sur un corpus de référence contenant approx. 4000 notices PASCAL relatives à la recherche en **optoélectronique** durant la période 1996-2003, et originellement divisé en 2 sous-périodes. => **4000 dimensions**

# ANALYSE DIACHRONIQUE DE LA RECHERCHE



## Principes mis en œuvre :

- ✓ Sous-périodes associées à différents points de vue,
- ✓ Optimisation du clustering en exploitant des méthodes neuronales à topologie libre (GNG) et des indices de qualité sans biais,
- ✓ Méthodes de caractérisation du contenu des classes basée sur l'étiquetage F-max,
- ✓ Raisonnement bayésien non supervisé adapté aux étiquettes.

# ANALYSE DIACHRONIQUE DE LA RECHERCHE

```

source cluster: 23 [19/10] target cluster: 2 [12/7]
- Stable labels - similarity kernel
f1: 0.259231[23] f2: 0.313356[ 8] Optical polymers (***)
f1: 0.086964[23] f2: 0.129486[ 2] Conducting polymers (***)

- Highly dominant (or peculiar) labels in source period
f1: 0.034510[23] f2: 0.000000[-1] Experimental study

- Highly dominant (or peculiar) labels in target period
f1: 0.072006[23] f2: 0.206426[ 2] Polymer films (***)
f1: 0.054435[23] f2: 0.114637[ 2] Polymer blends (***)
f1: 0.000000[-1] f2: 0.039558[ 2] Spin-on coating
f1: 0.000000[-1] f2: 0.028204[ 2] Polymerization
    
```

Théorie vers pratique

```

source cluster: 15 [22/9] target cluster: 24 [20/8]
- Stable labels - similarity kernel
f1: 0.038370[15] f2: 0.044230[24] Silicon compound (***)

- Highly dominant (or peculiar) labels in source period
f1: 0.043265[15] f2: 0.000000[-1] MIS structure
f1: 0.026522[15] f2: 0.000000[-1] Diamond

- Highly dominant (or peculiar) labels in target period
f1: 0.061132[15] f2: 0.222402[24] Amorphous semiconductors (***)
f1: 0.054647[15] f2: 0.131473[24] Hydrogen (***)
f1: 0.000000[-1] f2: 0.067403[24] Selenium
f1: 0.000000[-1] f2: 0.039028[24] Plasma CVD coatings
    
```

Nouveau composant

```

source cluster: 14 [18/6] target cluster: 14 [29/7]
- Stable labels - similarity kernel
f1: 0.035721[14] f2: 0.041813[14] Surface emitting laser (***)

- Highly dominant (or peculiar) labels in source period
f1: 0.148633[14] f2: 0.057783[14] Semiconductor laser (***)
f1: 0.078080[14] f2: 0.033436[14] Laser diodes (***)
f1: 0.026498[14] f2: 0.000000[-1] Surface
f1: 0.026027[14] f2: 0.000000[-1] Waveguide laser

- Highly dominant (or peculiar) labels in target period
f1: 0.000000[-1] f2: 0.068895[14] Light sources
f1: 0.000000[-1] f2: 0.039487[14] Laser beam applications
f1: 0.000000[-1] f2: 0.029637[14] Vertical cavity laser
f1: 0.000000[-1] f2: 0.025024[14] VCSEL
    
```

Théorie vers pratique

```

source cluster: 24 [23/9] target cluster: 33 [27/13]
- No stable labels

- Highly dominant (or peculiar) labels in source period
f1: 0.266901[24] f2: 0.068167[33] Optical fabrication (***)
f1: 0.045998[24] f2: 0.000000[-1] Integrated circuit technology
f1: 0.042258[24] f2: 0.000000[-1] Interference filter
f1: 0.041773[24] f2: 0.000000[-1] Semiconductor technology

- Highly dominant (or peculiar) labels in target period
f1: 0.077799[24] f2: 0.213749[33] Optical design techniques (***)
f1: 0.000000[-1] f2: 0.055834[33] Aberrations
f1: 0.000000[-1] f2: 0.000000[-1] Ray tracing
    
```

Changement de vocabulaire

```

source cluster 16 is vanishing
f1: 0.141849[16] f2: 0.000000[-1] Optical fiber
f1: 0.078762[16] f2: 0.000000[-1] Fiber laser
f1: 0.060706[16] f2: 0.000000[-1] Acoustooptical device
f1: 0.049628[16] f2: 0.000000[-1] Ring laser
    
```

```

target cluster 9 is appearing
f1: 0.035520[ 5] f2: 0.160462[ 9] Fluorescence
f1: 0.000000[-1] f2: 0.082686[ 9] Phosphorescence
f1: 0.063888[ 1] f2: 0.105132[ 9] Exciton
    
```

```

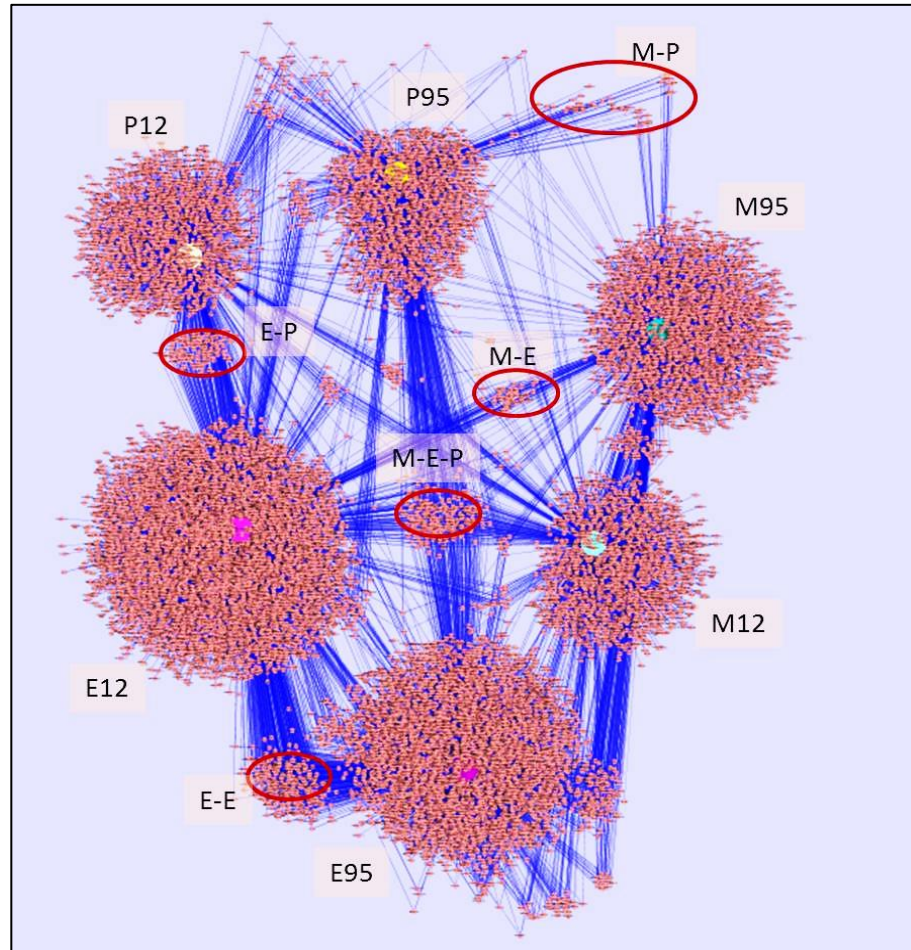
target cluster 39 is appearing
f1: 0.000000[-1] f2: 0.144184[39] Pixel
f1: 0.000000[-1] f2: 0.110076[39] CMOS image sensors
f1: 0.000000[-1] f2: 0.077578[39] Chip
f1: 0.000000[-1] f2: 0.060044[39] High sensitivity
    
```



# PERSPECTIVE : VISUALISATION

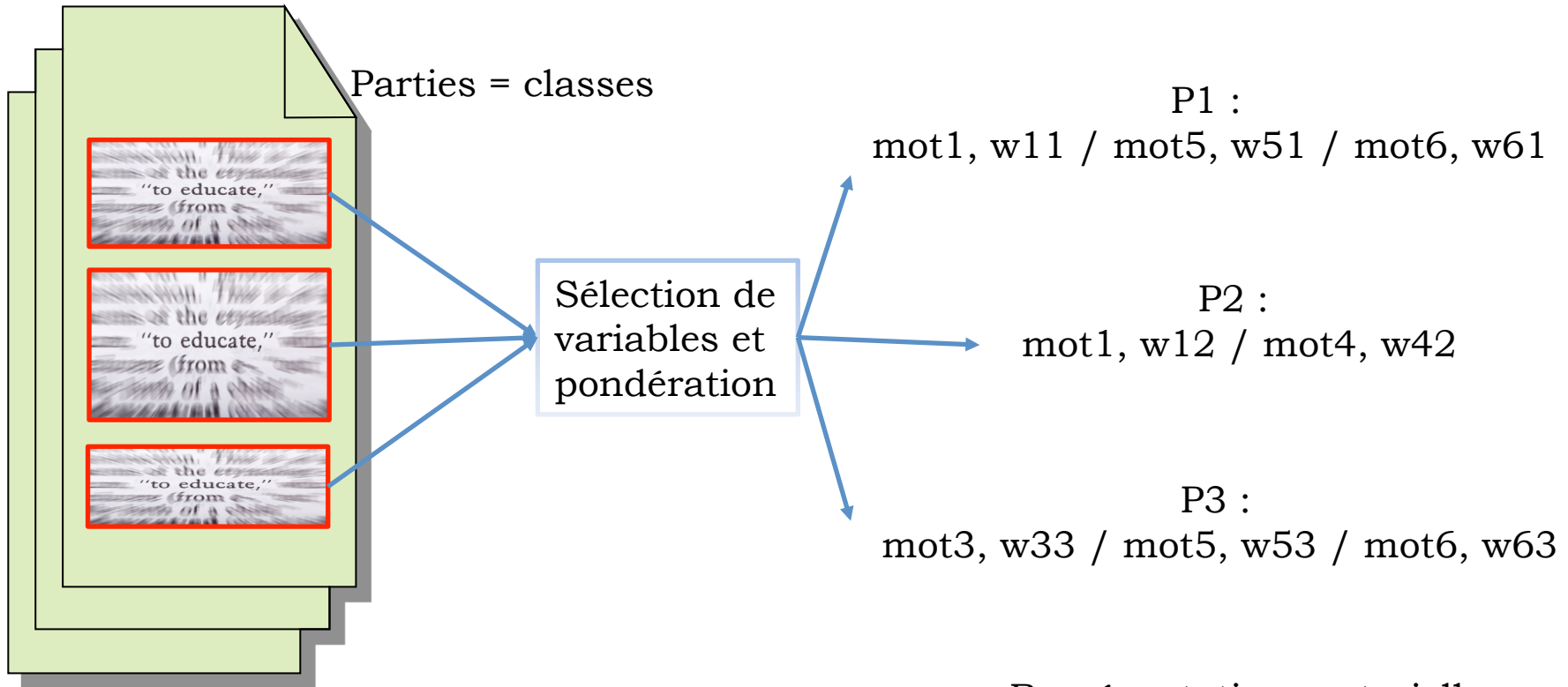
## Graphe Auteurs-Classes

[Cuxac et Lamirel, 2013]



- **Identification des passeurs de savoirs.**

# PERSPECTIVE : INDEXATION ET RÉSUMÉ AUTOMATIQUE



Article  
structuré  
=  
Classification

F sélection + contraste

Représentation vectorielle :  
 $D = w1.1, 0, w3.3, w4.2, w5.3, w6.1$

Représentation graphique :





## Article choisi au hasard dans les réservoirs Istex

# Subsequent insect stings in children with hypersensitivity to Hymenoptera

Pia Hauk, MD, Katrin Friedl, Klaus Kaufmehl, MD, Radvan Urbanek, MD, and Johannes Forster, MD

From University Children's Hospitals, Freiburg, Germany, and Vienna, Austria

**To investigate the risk of life-threatening reactions to future stings, we sequentially challenged 113 children (aged 2 to 17 years) allergic to insect stings with a sting by the relevant insect. The time interval between the challenges varied from 2 to 6 weeks. The history of the index stings was a large local reaction (LR) in 16% and a systemic reaction (SR) in 84% of the test subjects. On the first challenge, 76% had a normal LR, 11% a large LR, and 13% an SR. On the second challenge, 78% of the children had a normal LR, 5% a large LR, and 17% an SR. Thirty-nine of the untreated children were exposed to a field sting during the subsequent 3-year follow-up period. In comparison with other diagnostic evaluations such as skin-prick tests, determinations of specific IgE and IgG antibodies, and single-sting exposure, the dual sting challenge scheme appears to be the best predictor of reactions to subsequent stings. It also appears to be helpful in selecting patients with an uncertain sensitization status for venom immunotherapy. (J PEDIATR 1995;126:185-90)**

In childhood, allergy to Hymenoptera venom is mainly caused by stings of honeybees and wasps. In Europe, yellow jackets are known as "wasps," whereas in the United States, Polistes wasps are known as "wasps."<sup>1</sup> Between 0.4% and 4% of the population have systemic allergic reactions to insect stings.<sup>2-4</sup> The incidence of systemic reactions to subsequent stings is lower in children and adolescents than in adults.<sup>5-8</sup> Prospective observations of the natural course of insect allergy show that adults have a risk of 27% to 57%<sup>3, 9-11</sup> of having repeated systemic allergic reactions, in comparison with a risk of 10% to 20% in children.<sup>4-6, 8</sup> Therefore venom immunotherapy should be indicated less frequently in children.<sup>8</sup> In vitro assays and risk scores provide only limited help in identifying those patients at risk of having further life-threatening allergic reactions. Numerous studies<sup>12-15</sup> have been unsuccessful in showing a correlation between the standard diagnostic methods—mainly skin-prick tests and measurements of specific IgE and IgG

antibodies—and the reactions to subsequent insect stings. Treatment recommendations based only on those criteria typically lead to an overestimation of the number of children who require venom immunotherapy.<sup>6, 8, 16</sup>

Although single diagnostic sting challenges give additional information, there is increasing concern about the possible booster effect. From the natural history of bee venom allergy, we know that one sting followed by another

See commentary, p. 257.

AU	Arbitrary unit(s)
LR	Local reaction
SR	Systemic reaction

2 to 4 weeks later will result in the highest incidence of systemic reactions. We tried to mimic this naturally occurring event by subjecting test subjects to sequential sting challenges to detect the group of patients at highest risk. Those who did not react and therefore were not assigned to receive venom immunotherapy were followed for up to 3 years for life-threatening events after natural stings.

Submitted for publication April 15, 1994; accepted Aug. 10, 1994.  
Reprint requests: Johannes Forster, MD, University Children's Hospital, Mathildenstr. 1, D-79106 Freiburg, Germany.  
Copyright © 1995 by Mosby-Year Book, Inc.  
0022-3476/95/\$3.00 + 0 9/20/59779

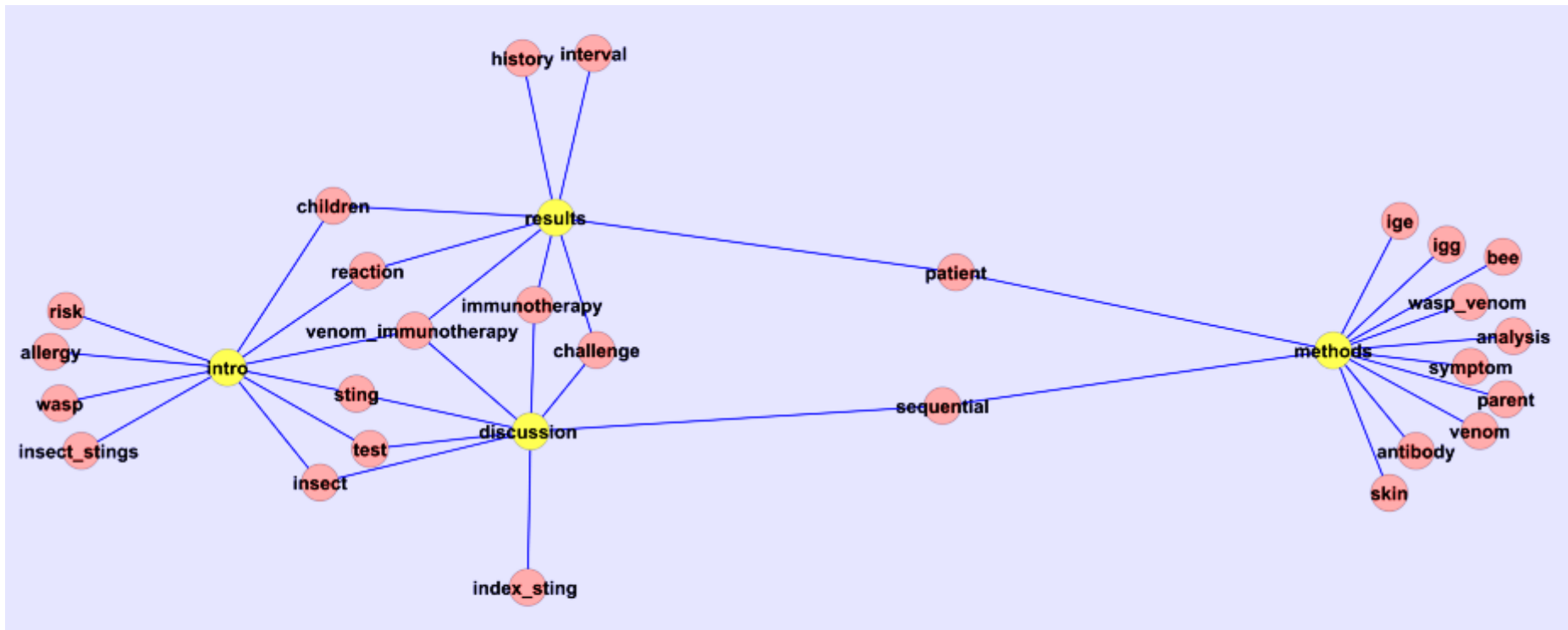
```

1<?xml version="1.0" encoding="utf-8"?><!DOCTYPE article PUBLIC "-//ES//DTD
2<ENTITY gr1 SYSTEM "gr1.NDATA IMAGE">[E
3<article docsubtype="fla" xml:lang="en">[E
4<item-info>[E
5<jid>YMPD</jid>[E
6<aid>59779</aid>[E
7<ce:pii>S0022-3476(95)70543-0</ce:pii><ce:doi>10.1016/S0022-3476(95)70543-0
8<ce:copyright type="full-transfer" year="1995">Mosby, Inc.</ce:copyright><
9<ce:floats><ce:table id="tab1" colsep="0" rowsep="0" frame="topbot"><ce:tal
recommenda
10<colspec colname="col2" colsep="0" /></colspec>
11<colspec colname="col3" colsep="0" /></colspec>
12<thead><row rowsep="1"><entry valign="bottom" colsep="1"></entry>[E
13<entry name="col2" nameend="col3" align="center" valign="bottom">Risk po
14</row>[E
15<row rowsep="1"><entry align="center" valign="bottom" colsep="1"></entry>[E
16<entry align="center" colsep="1" valign="bottom" colsep="1"></entry>[E
17<entry align="center">2</entry>[E
18</row>[E
19</thead>[E
20<tbody>[E
21<row rowsep="1"><entry colsep="1">Reaction to index sting</entry>[E
22<entry align="center" colsep="1">Large LR, moderate SR</entry>[E
23<entry align="center">Severe SR</entry>[E
24</row>[E
25<row rowsep="1"><entry colsep="1">Skin-prick test (100 &#x03BC;g/ml venom)
26<entry align="center" colsep="1">wheal diameter &#x003C;3 mm</entry>[E
27<entry align="center">wheal diameter &#x2265;3 mm</entry>[E
28</row>[E
29<row rowsep="1"><entry colsep="1">Specific IgE (RAST)</entry>[E
30<entry align="center" colsep="1">Class 1</entry>[E
31<entry align="center">Classes 2-4</entry>[E
32</row>[E
33<row><entry name="col1" nameend="col3"></entry>[E
34</row>[E
35</tbody></tgroup>[E
36<ce:legend><ce:simplepara><ce:italic>RAST</ce:italic>, radioallergosorbent
</ce:simplepara></ce:legend><ce:table id="tab2" colsep="0" row
</ce:simplepara><ce:caption><tgroup cols="5"><colspec colname="col1" col
37<colspec colname="col2" colsep="0" /></colspec>
38<colspec colname="col3" colsep="0" /></colspec>
39<colspec colname="col4" colsep="0" /></colspec>
40<colspec colname="col5" colsep="0" /></colspec>
41<thead><row rowsep="1" valign="bottom"><entry colsep="1">Index sting react
42<entry align="center" colsep="1">n</entry>[E
43<entry align="center" colsep="1">Age (yr) median: range</entry>[E
44<entry align="center" colsep="1">Male/Female</entry>[E
45<entry align="center" colsep="1">Risk score &#x2265;5 points</entry>[E
46</row>[E
47</thead>[E
48<tbody>[E
49<row rowsep="1"><entry colsep="1">Large LR</entry>[E
50<entry align="center" colsep="1">18</entry>[E
51<entry align="center" colsep="1">9; 2-17</entry>[E
52<entry align="center" colsep="1">12/6</entry>[E
53<entry align="center">4</entry>[E
54</row>[E
55<row rowsep="1"><entry colsep="1">Moderate SR</entry>[E
56<entry align="center" colsep="1">57</entry>[E
57<entry align="center" colsep="1">8; 3-16</entry>[E
58<entry align="center" colsep="1">33/24</entry>[E
59<entry align="center" colsep="1">33/24</entry>[E

```

# PERSPECTIVE : INDEXATION ET RÉSUMÉ AUTOMATIQUE

Subsequent insect stings in children with hypersensitivity to Hymenoptera



Représentation graphique du full-text  
Représentation vectorielle du full-text (→ clustering...)

# PERSPECTIVE : INDEXATION ET RÉSUMÉ AUTOMATIQUE

Génération automatique de résumé par  
extraction de phrases

## Subsequent insect stings in children with hypersensitivity to Hymenoptera

Although there were more severe reactions in the group of children who required immunotherapy according to our assessment, no significant correlation could be detected between the reactions to the index sting and to the challenge stings, or between the reactions to the index sting and to the field sting. 98.2682

Considering the previous reaction to the index sting and the results of skin-prick tests and venom-specific IgE measurements as criteria for the recommendation of venom immunotherapy, 41% of the scored bee venom- and wasp venom-allergic children would have been assigned to this treatment, but only 9% received venom immunotherapy as a result of the clinical reaction to the second challenge. 98.6776

Although this is not a 100% safety record, we believe that the sequential insect sting challenge performed in the hospital represents the safest and most informative method of eliminating unnecessary venom immunotherapy in children having mild to moderate SRs to an index sting. 137.1604

On the basis of the data presented, we suggest the following diagnostic and therapeutic procedures for children up to 16 years of age: Sensitized patients, identified by a positive skin-prick test result or specific IgE finding, who had only a large LR to the index sting, need neither a challenge sting nor venom immunotherapy. 104.307

# CONCLUSIONS ET PERSPECTIVES

- Nouvelle approche statistique pour l'analyse des textes basée sur la maximisation de l'étiquetage.
- Cette approche répond aux contraintes liées au traitement des informations textuelles en ligne, volumineuses, changeantes et/ou déséquilibrées.
- Elle s'applique à l'analyse supervisée tout comme à l'analyse non supervisée incrémentales.
- Nombreuses applications potentielles dans le domaine :
  - Stylométrie, Analyse du plagiat,
  - Construction de lexiques, ontologies,
  - Classification automatique des textes,
  - Analyses des réseaux d'auteurs et de leur interaction,
  - Analyse diachronique et analyse des flux d'information textuelle (Projet ISTEEX-R).

# CONCLUSIONS ET PERSPECTIVES

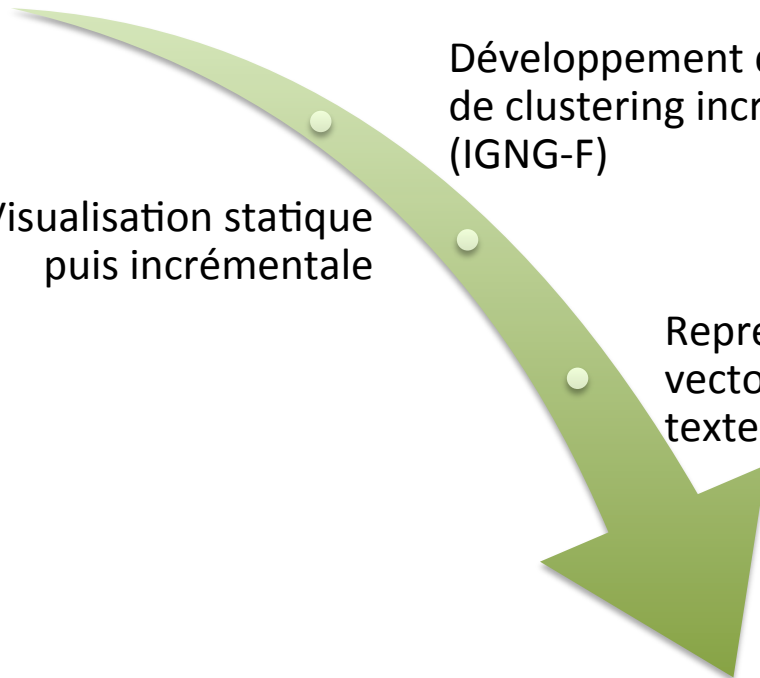
Développement de  
la méthode  
diachronique.

Développement de la méthode  
de clustering incrémental  
(IGNG-F)

Visualisation statique  
puis incrémentale

Représentation  
vectorielle du  
texte plein

Visualisation des résultats et  
interactions avec l'utilisateur



# Conclusion

- Des méthodes de fouille de données sur des objets complexes extraites de textes
- Des méthodes de TAL robustes qui peuvent transformer le texte en données
- Une méthode itérative et interactive
  
- Quelques problèmes :
  - Les anaphores
  - Les corpus pluri-domaines
  - La robustesse dans le traitement du volume des données : compromis précision/quantité