Donner accès au contenu des documents textuels

Acquisition de connaissances et analyse de corpus spécialisés

Habilitation à diriger les recherches

Adeline Nazarenko

10 décembre 2004

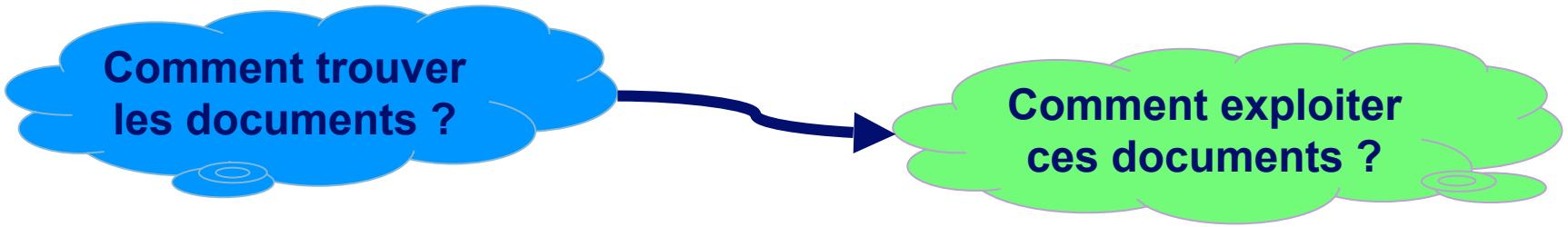
LIPN - UMR 7030
Université Paris13

Plan

1. **Introduction**
   - Problématique
   - Contexte
2. **Acquisition de connaissances**
   - Objectif et démarche
   - Acquisition de termes
   - Mise en réseau des termes
   - Bilan
3. **Analyse de documents**
   - Un problème d'extraction d'information
   - Une démarche : combiner TAL et Apprentissage
   - Une méthode de normalisation de textes
4. **Conclusion et perspectives**

ACCÈS AU CONTENU DES DOCUMENTS

Au-delà des techniques de recherche d'information,
Accéder au contenu des documents



Le contenu

vue globale → *éléments précis*

Recherche d'information dans une base documentaire

Visualisation de documents

Résumé automatique

Outils de navigation

Indexation fine de documents

Hypertextualisation

Questions-réponses

Extraction d'information

LE POINT DE DÉPART
Le projet Kalipsos

Contexte

- Années 1988-94
- Centre Scientifique d'IBM France (J. Fargues)
- Travail de thèse (sept 1990-janv 1994)

Un projet ambitieux

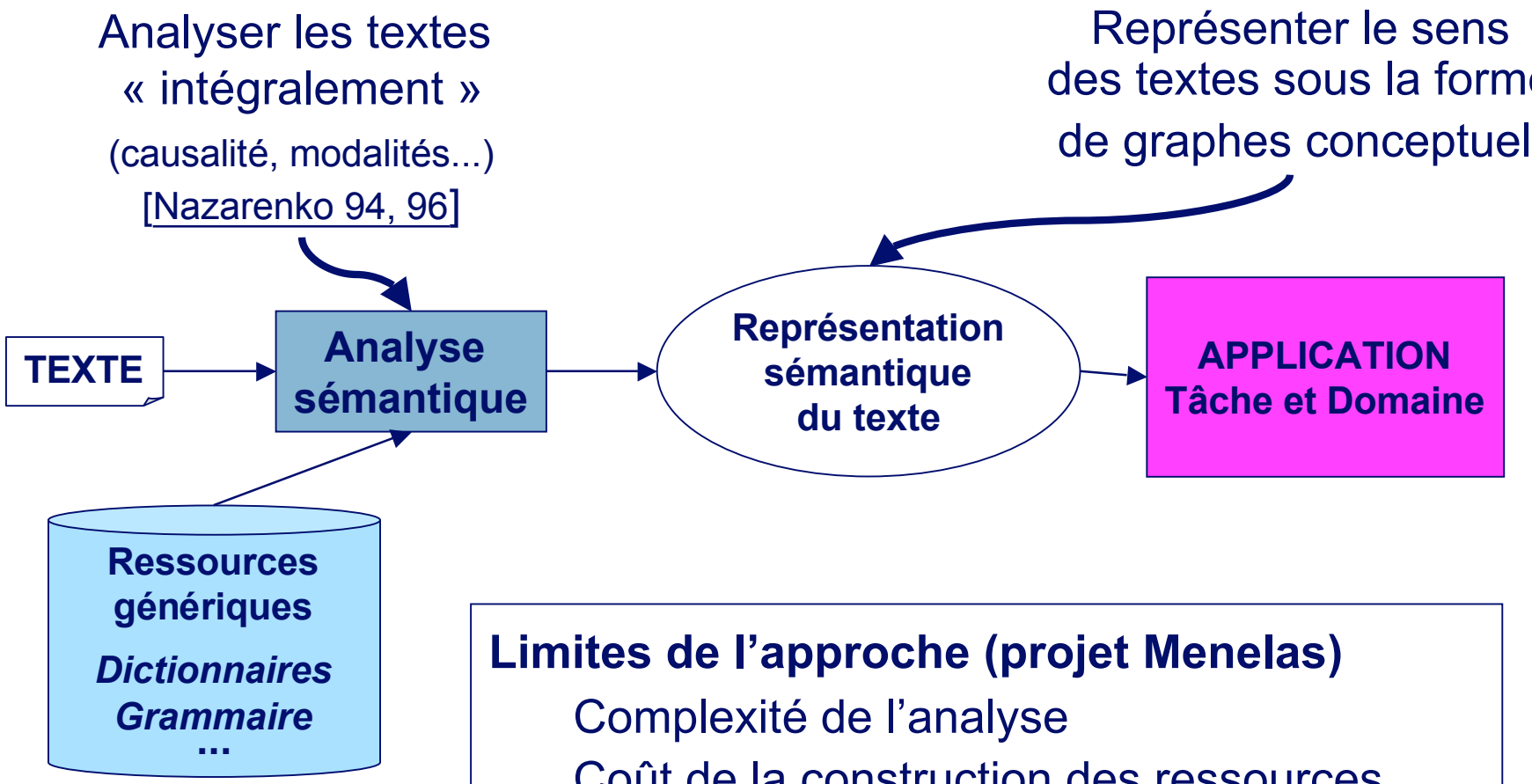
Elaborer un système *générique* de compréhension de texte

- applicable à tout type de texte
- utilisable pour différentes applications

Questions-réponses **Filtrage de dépêches de presse**

Analyse de compte-rendus médicaux

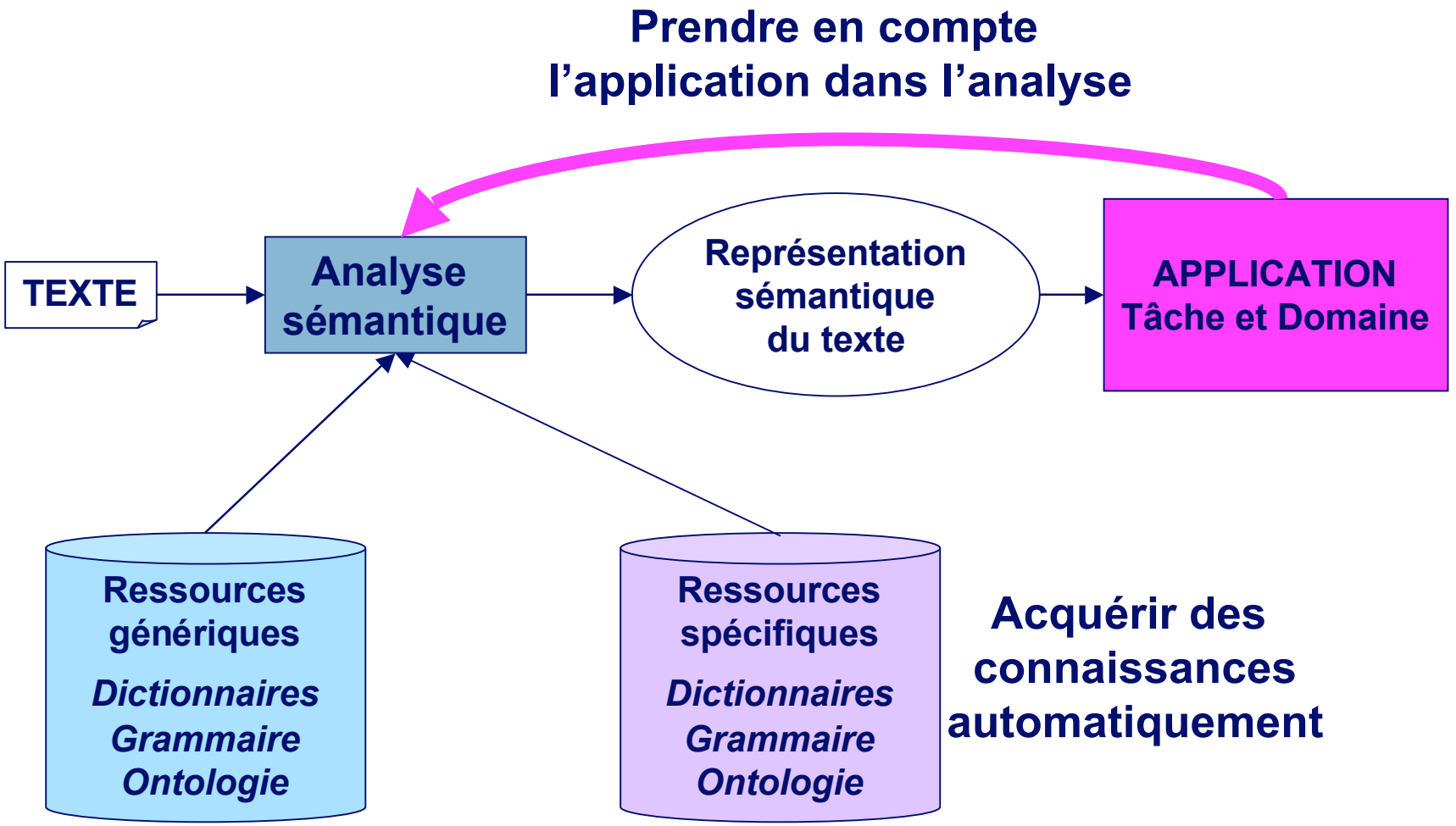
L'APPROCHE KALISPOS



Limites de l'approche (projet Menelas)

- Complexité de l'analyse
- Coût de la construction des ressources
- Insuffisante prise en compte de l'application
- Evaluation problématique

DOUBLE OBJECTIF



TAL et RC : DES DOMAINES EN ÉVOLUTION

Mutations du TAL

- Emergence de la notion de corpus [Habert et al. 97]
- De nouveaux objectifs : efficacité et couverture
- L'essor des méthodes statistiques
- Un certain renouveau de l'approche sémantique

sémantique référentielle ➜ **sémantique distributionnelle et textuelle**

Renouveau de l'analyse terminologique [Meyer et al. 92](Groupe TIA)

terme comme étiquette de concept ➜ **terme comme unité textuelle**

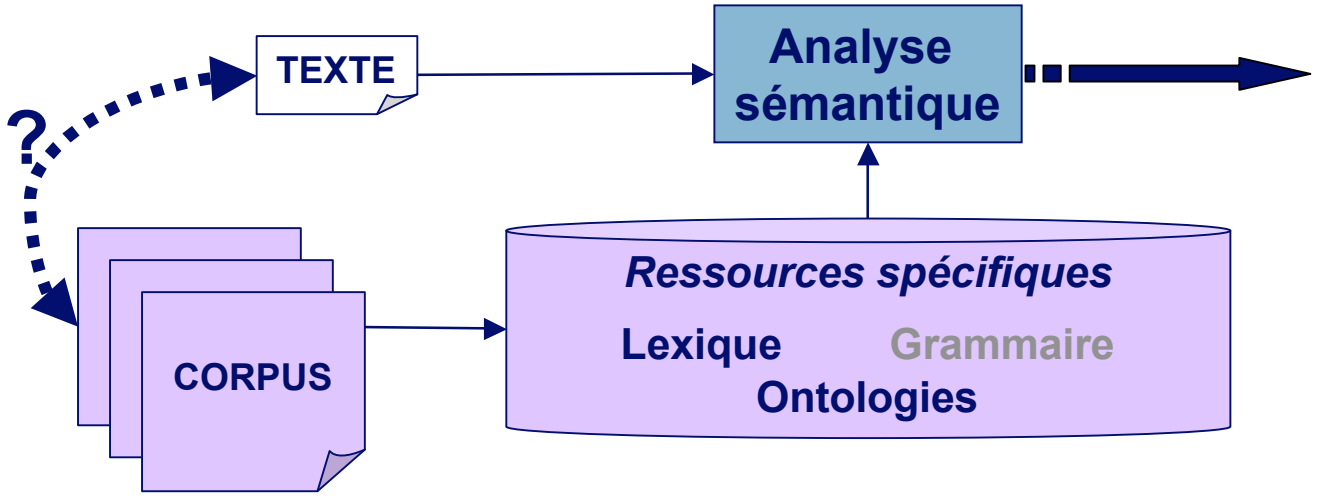
Attention portée aux ontologies

- Multiples projets de construction d'ontologies : de Cyc à KA2
- Réflexion sur les propriétés formelles [Guarino 95]
- Définition de nouveaux langages (défi du web sémantique)

Plan

1. **Introduction**
   - Problématique
   - Contexte
2. **Acquisition de connaissances**
   - Objectif
   - Acquisition de termes
   - Mise en réseau des termes
   - Bilan
3. **Analyse de documents**
   - Un problème d'extraction d'information
   - Une démarche : combiner TAL et Apprentissage
   - Une méthode de normalisation de textes
4. **Conclusion et perspectives**

ACQUÉRIR DES CONNAISSANCES
À partir de corpus

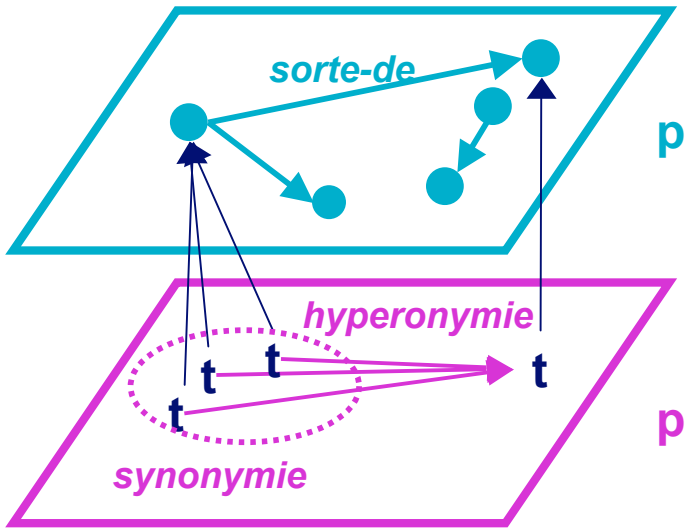


- ✓ Matériau disponible
- ✓ Articulation des niveaux linguistique et ontologique [Nédellec & Nazarenko 04]

- ✖ Choix difficile des corpus à exploiter
- ✖ Biais linguistique de l'approche [Bouaud et al. 00]

QUE FAUT-IL ACQUÉRIR ?
Un embryon d'ontologie

- ✓ Termes, unités sémantiques qui renvoient aux concepts du domaine
- ✓ Relations entre ces termes, reflet partiel des relations entre concepts
- ✖ Ontologie formalisée [Biébow & Szulman 00]



plan conceptuel

plan linguistique

ACQUISITION DE TERMES

Quel est le vocabulaire du domaine ?

`transcriptional factor` `DNA replication`

`sporulation-specific transcription factor sigma G`

1ère approche : Acquisition en corpus

[Bourigault 94, Justeson & Katz 95, Daille 96...]

Expériences diverses

Extracteurs : Lexter, Syntex, Acabit

Domaines : médecine, génomique

Objectifs : construction d'index, extraction d'information

➢ Des résultats intéressant mais bruités

2ème approche : Projection de ressources existantes

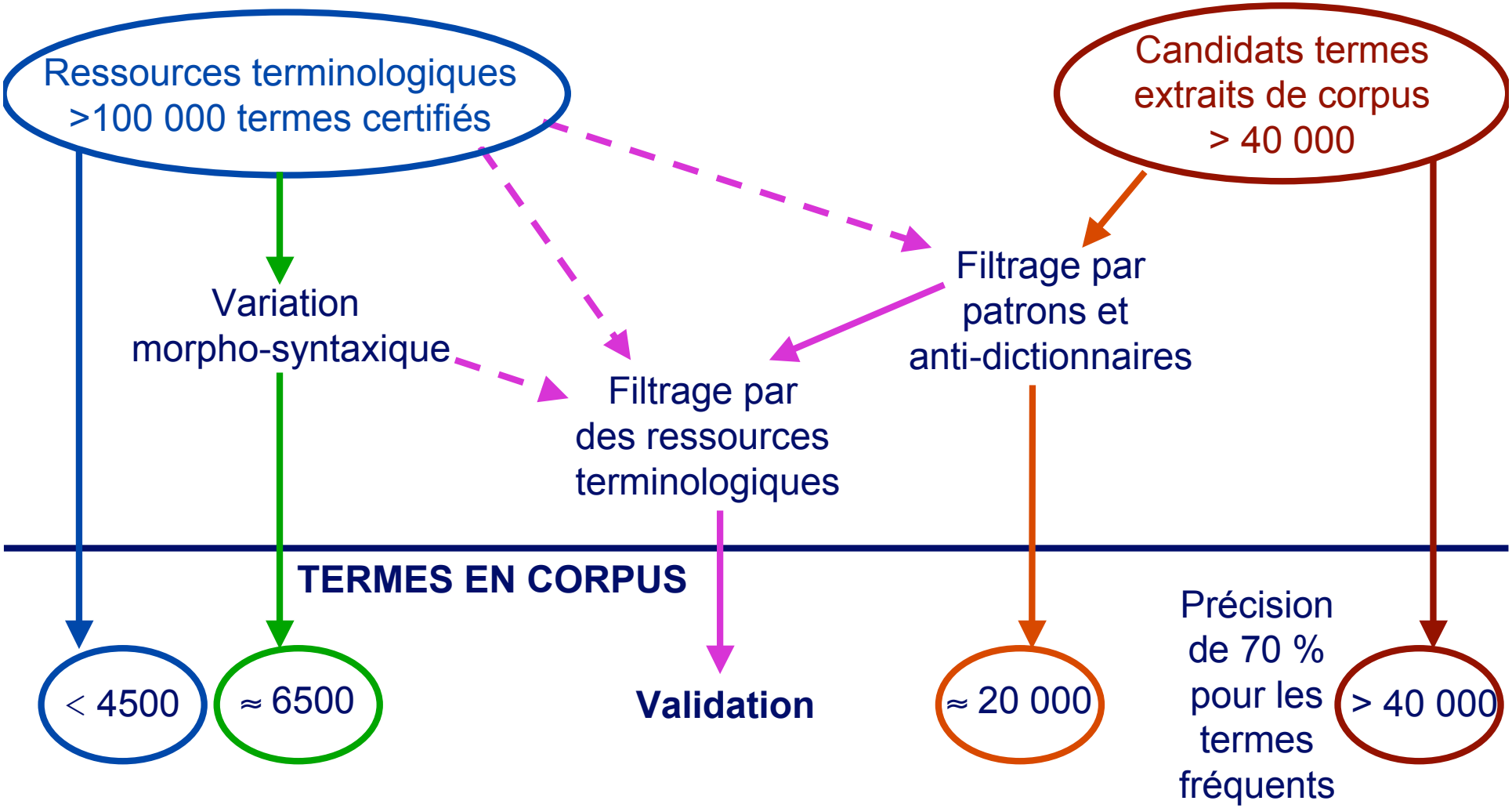
Expérience en génomique : exploitation de différentes ressources spécialisées

➢ Comment choisir les ressources ?

➢ Comment pallier le manque de couverture des ressources ?

COMBINAISON DES DEUX APPROCHES

[Hamon 04]



CONSTRUIRE UN RÉSEAU DE TERMES
Des travaux épars

Différents types de relations : *hyperonymie, **synonymie**, est-le-symptôme-de ...*
[Hamon *et al*. 98]

Différentes approches : **structurelle, contextuelle, distributionnelle**...

Expériences sur des corpus divers

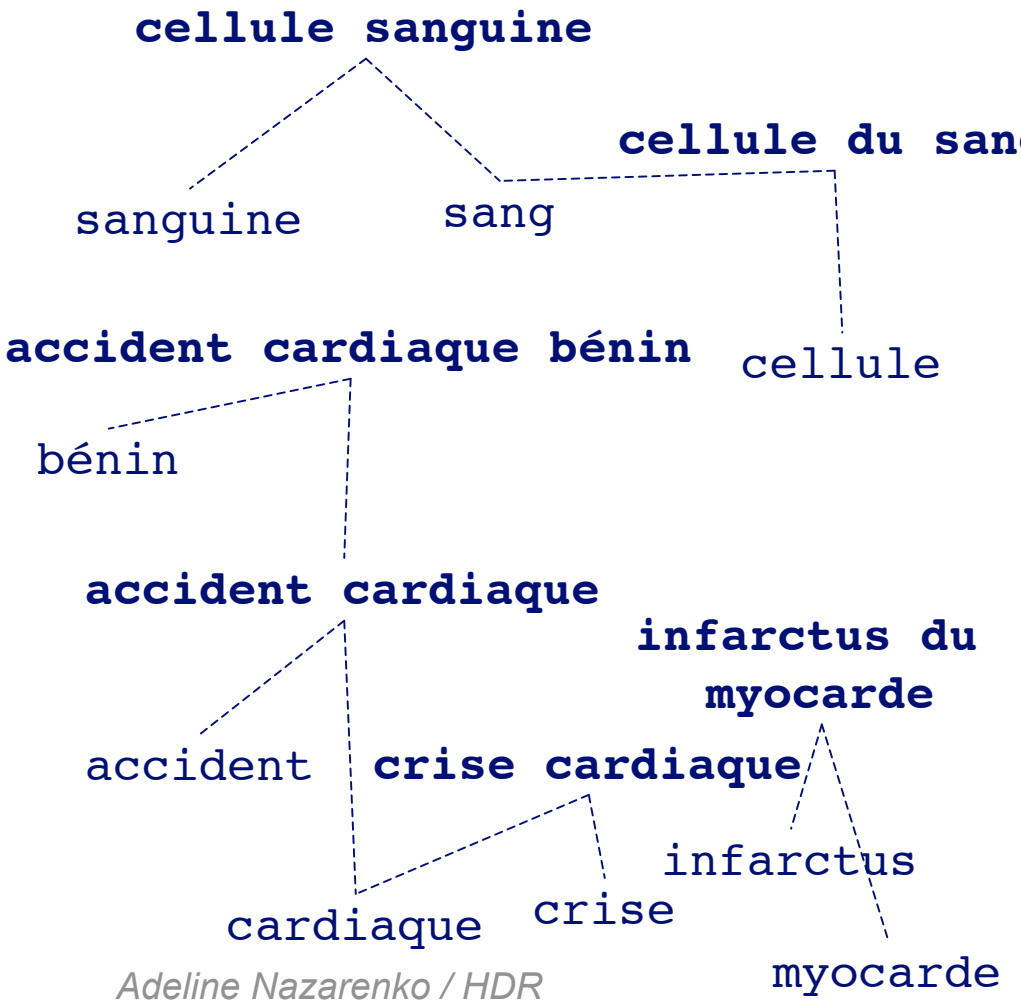
Différents types de corpus :	**corpus spécialisés (médecine, génomique)**
Corpus de tailles variées	**< 500 Kmots**
Corpus de différentes langues	**anglais, français**

Objectifs divers

- Enrichissement de thesaurus [Yarowsky 92, Morin 99]
- Normalisation terminologique [Dagan & Church 96, Hamon & Nazarenko 01, ...]
- Construction d'index [Aït El Mekki & Nazarenko 03]

Une travail d'intégration

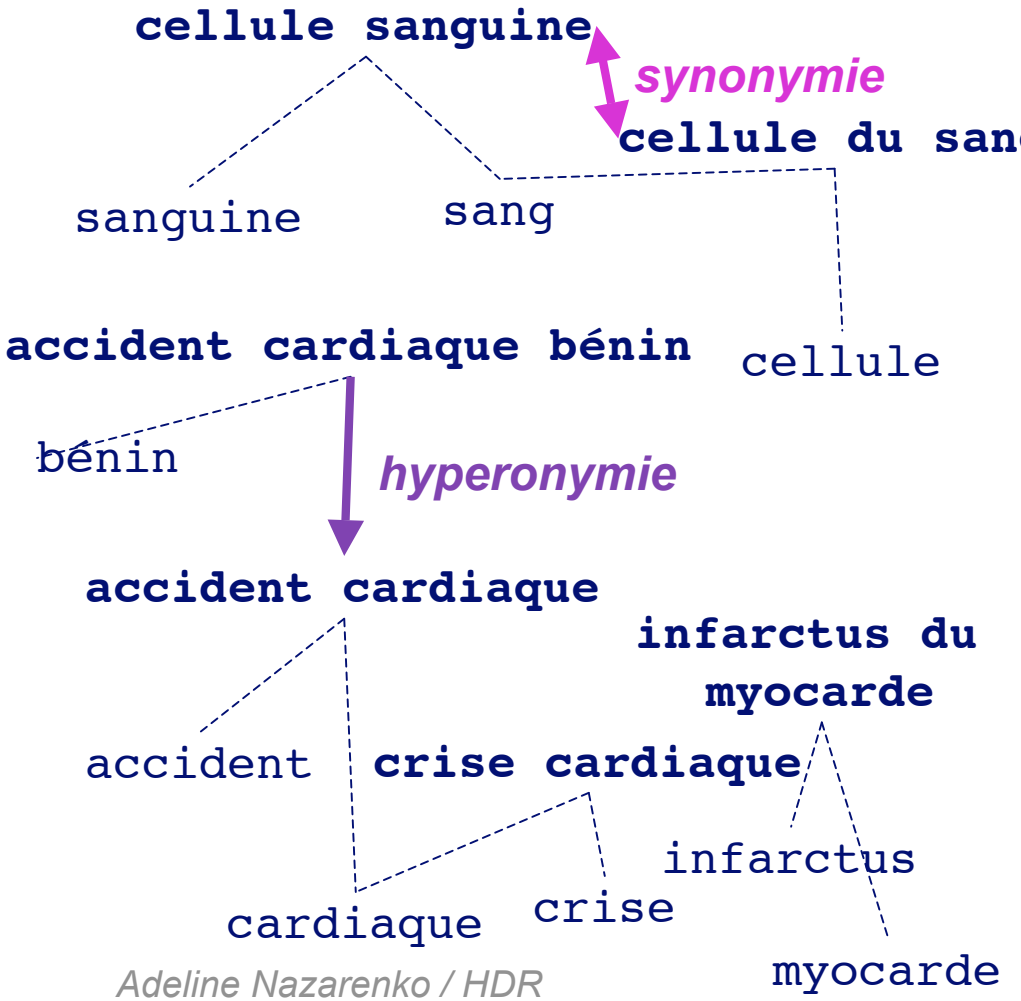
CONSTRUIRE UN RÉSEAU DE TERMES
Réseau syntaxique



Liste de termes

Liens de composition syntaxique [Bourigault 94]

CONSTRUIRE UN RÉSEAU DE TERMES
Règles structurelles



Règles de composition syntaxique [Dagan & Church 96]

→

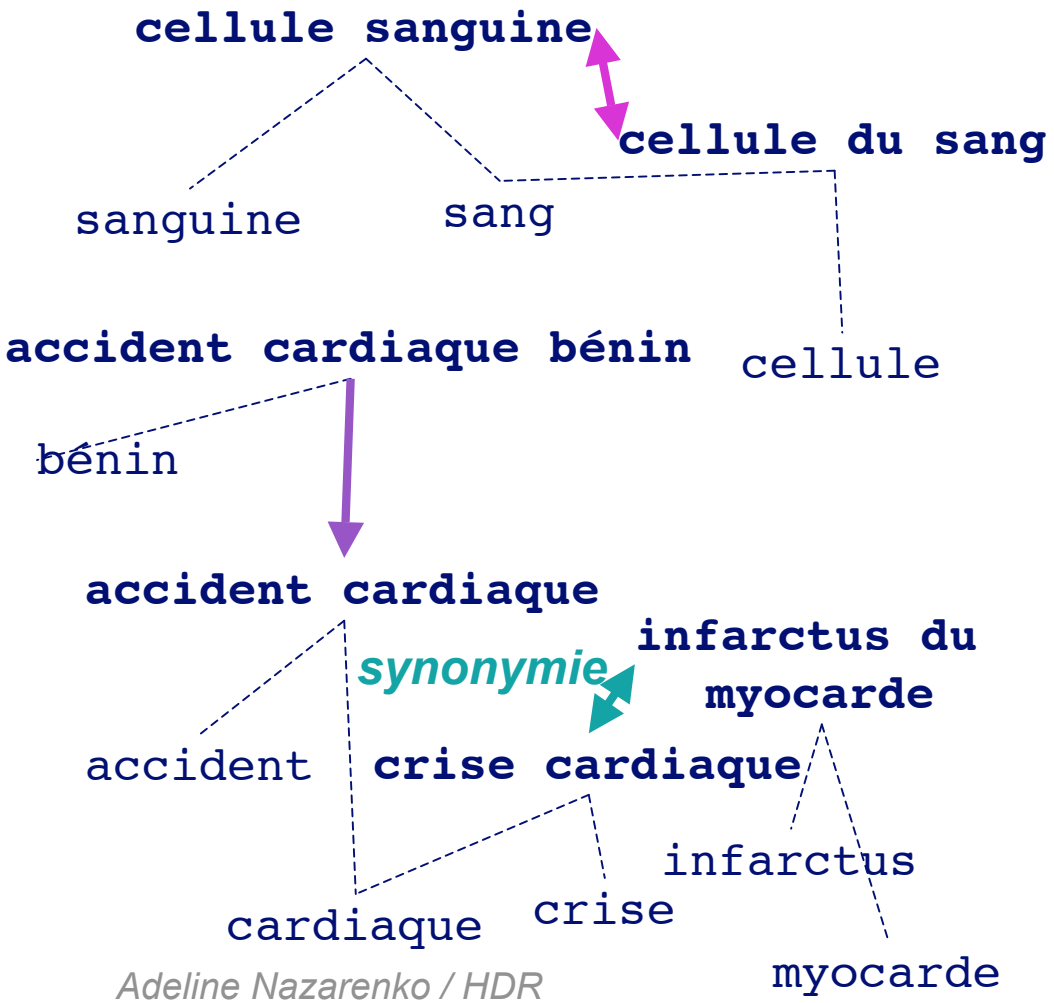
Règles de variation morpho-syntaxique [Jacquemin 97]

↔

Méthode
- endogène
- assez fiable
- très productive

Relations peu originales

CONSTRUIRE UN RÉSEAU DE TERMES
Règles contextuelles



Hyperonymie
[Hearst 92, Morin99, ...]

Synonymie ⟷
[Pearson 98, Weissenbacher 04]

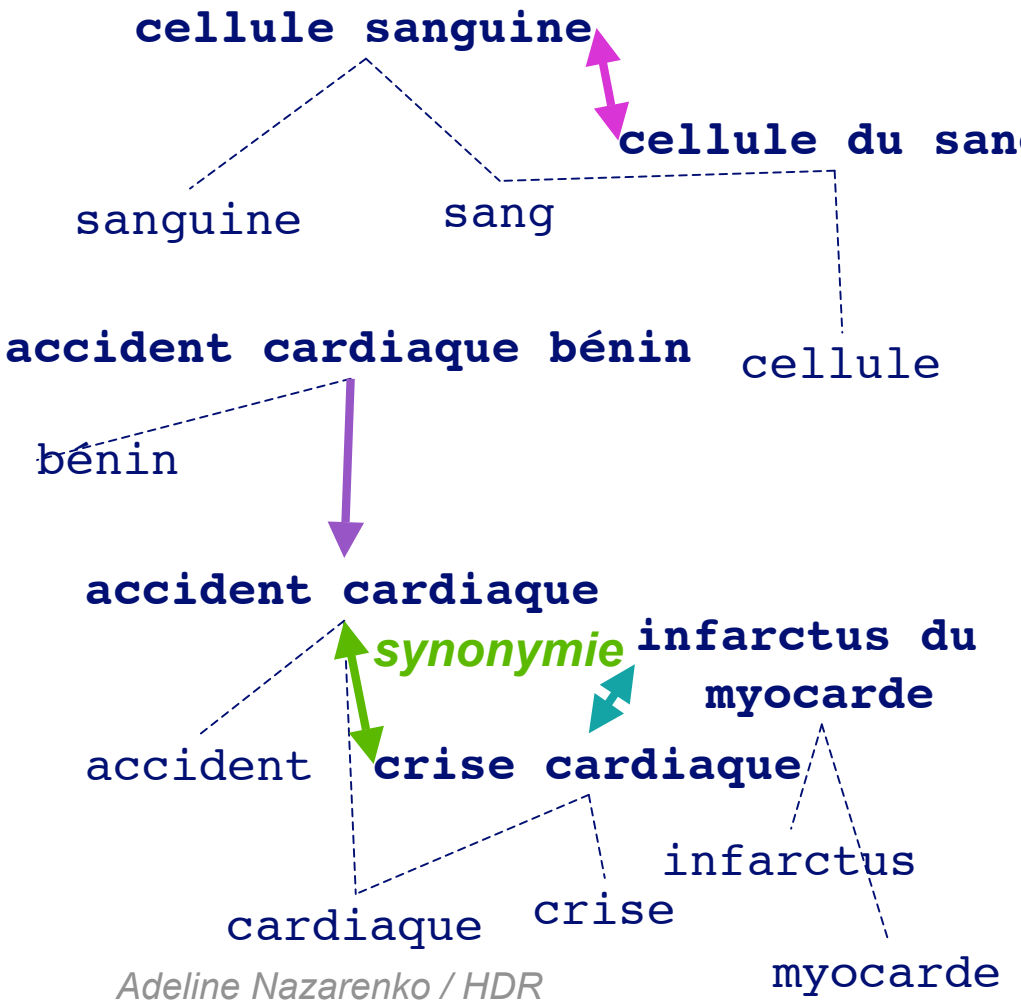
N1 *(ou* N2*)*
*crise cardiaque **(ou** infarctus du myocarde**)***

N1*, également appelé* N2*,*
N1 *, anciennement* N2*,*

Méthode
- fiable
- peu productive

Relations originales

CONSTRUIRE UN RÉSEAU DE TERMES
Exploitation de connaissances lexicales



Synonymie ⟷

[Hamon & Nazarenko 01]

crise=accident

crise cardiaque≈accident cardiaque

crise d'asthme≠accident d'asthme

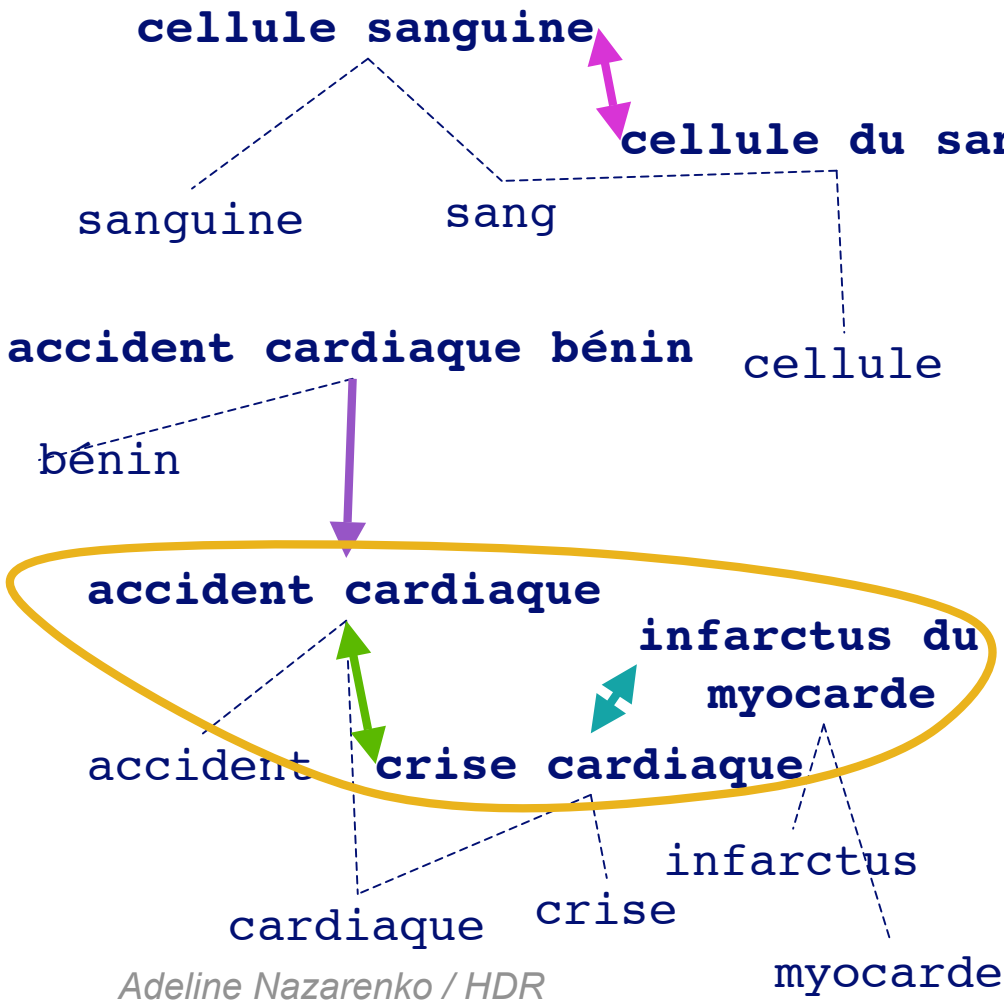
Hyperonymie

Méthode

- moyennement fiable
- peu productive

Relations originales

CONSTRUIRE UN RÉSEAU DE TERMES
Sémantique distributionnelle

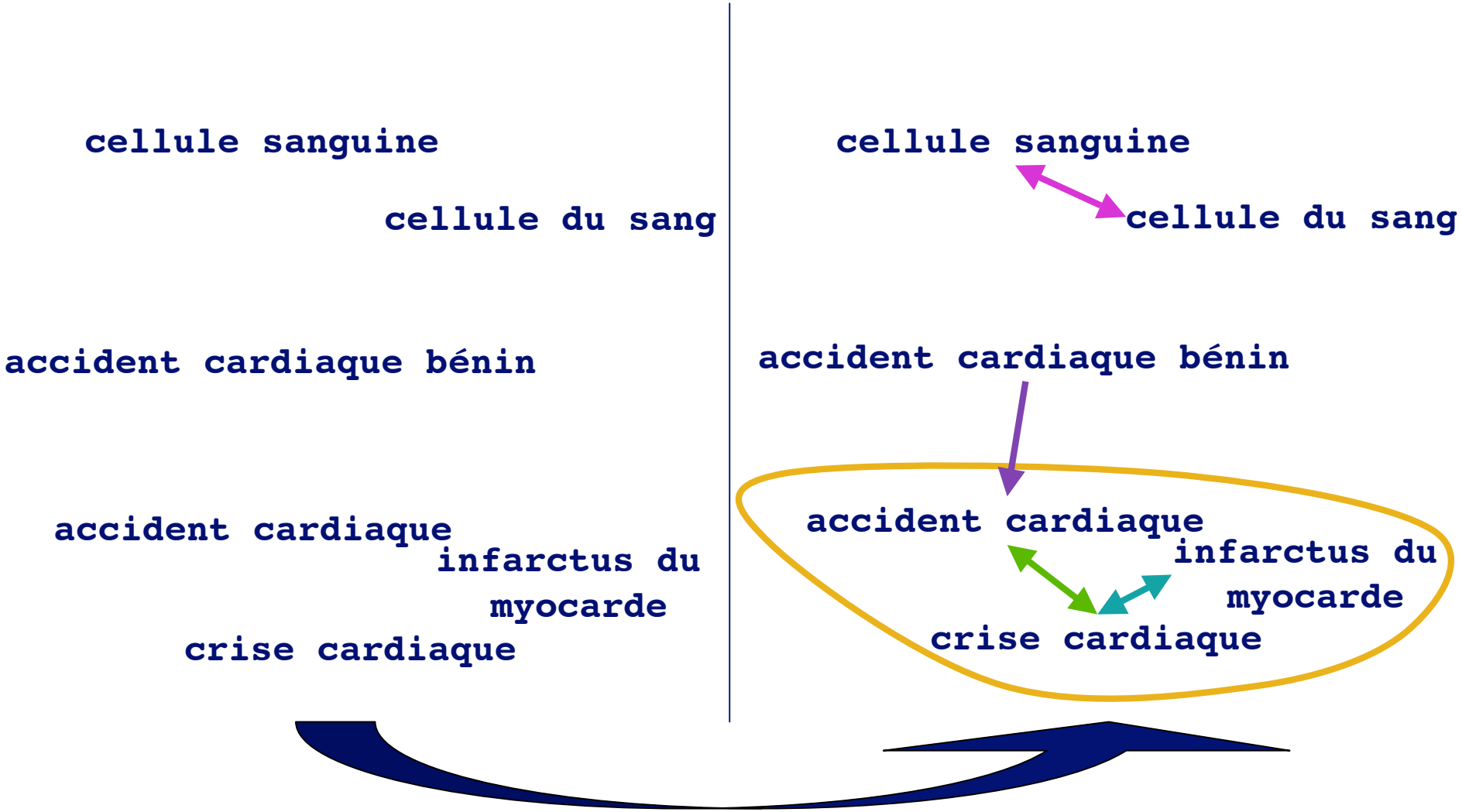


Proximité sémantique
[Mac Mahon & Smith 96, Habert & Nazarenko 96, Faure & Nédellec 99, Dagan et al. 99, ...]

Petits corpus
- Partir de contextes syntaxiques
- Prendre en compte tous les termes

➢ Résultats intéressants mais bruités

CONSTRUIRE UN RÉSEAU DE TERMES
De la liste au réseau



CONSTRUIRE UN RÉSEAU DE TERMES
Bilan

[Aït El Mekki & Nazarenko 05]

Intégrer les différentes méthodes

1. Bâtir l'ossature du réseau	←	Règles structurelles
2. Ajouter des relations	←	Règles contextuelles Exploitation de connaissances lexicales
3. Saturer du réseau	←	Règles de propagation [Hamon 02]
4. Pondérer les résultats	←	Sémantique distributionnelle

Gagner en robustesse

- Une méthode n'est pas également productive sur tous les corpus
- Exemple : repérage contextuel de la synonymie

✖ entre termes (corpus Menelas)

✓ entre noms de gènes (corpus Caderige)

DÉMARCHE GLOBALE D'ACQUISITION

1. **Choix des ressources**
   Se doter de métriques permettant d'apprécier l'adéquation d'une ressources à un corpus [Ninova 04]

2. **Extraction de termes**

3. **Structuration en intégrant les différentes méthodes**

4. **[Validation]** [Hamon 02, Aït El Mekki 04]
   - Développer des interfaces
     Vue globale sur le réseau, retour au texte...
   - Propager les résultats de validation dans le réseau
     Apprentissage incrémental?
   - Trier les résultats en fonction l'application
     Fréquence des termes
     + Autres critères (discursifs, sémantiques...)

ACQUISITION DE CONNAISSANCES
Défi : construire une expertise

Une méthode à paramétrer différemment pour chaque application ?

- Le genre de corpus
- Les ressources disponibles
- L'effort d'acquisition envisageable
- Le volume de données
- La qualité d'analyse visée

Les expériences sont coûteuses

- Préparation des données
- Validation
- Interprétation des résultats

➢ **Il faut développer des outils et méthodes de test**

- Atelier Mo'K pour tester des distances sémantiques [Bisson et al. 00]
- Mesures de prédiction de l'apprentissage de noms de personnes [Poibeau 02]
- Mesures de prédiction de l'adéquation des ressources aux corpus [Ninova 04]

➢ **Poursuivre dans cette voie**

Plan

1. **Introduction**
   - Problématique
   - Contexte
2. **Acquisition de connaissances**
   - Objectif et démarche
   - Acquisition de termes
   - Mise en réseau des termes
   - Bilan
3. **Analyse de documents**
   - Un problème d'extraction d'information
   - Une démarche : combiner TAL et Apprentissage
   - Une méthode de normalisation de textes
4. **Conclusion et perspectives**

ANALYSE DE DOCUMENTS

Une analyse sémantique modeste

Représenter le sens des phrases, du texte →

1. **Repérer certaines unités textuelles saillantes sur le plan sémantique**
2. **Les typer**
3. **Eventuellement les mettre en relation**

Une analyse sémantique paramétrable en fonction de l'application

- En intégrant des ressources acquises à partir de corpus
- En énonçant des règles d'interprétation

Trois directions de travail

- Extraction d'information (1998-2004)
- Création d'index de documents (2001-2004)
  Logiciel IndDoc (Thèse de T. Aït El Mekki)
- Recherche d'information spécialisée (2004-06)
  Projet européen STREP Alvis

EXTRACTION D'INFORMATION DANS LES TEXTES

Extraire des informations factuelles précises d'un ensemble de documents homogènes pour remplir automatiquement un formulaire défini à l'avance

Conférences MUC 1987-98

[Cowie & Lehnert 96, Grishman & Sundheim 96, Pazienza 97]

Une analyse locale

- Repérage d'un mot clef
- Analyse de son contexte

Une analyse guidée par des « règles d'extraction » de type SI... ALORS...

SI le fragment de texte considéré vérifie certaines conditions (prémisse de la règle)

ALORS on peut remplir tout ou partie du formulaire (conclusion de la règle)

Une problématique relativement peu représentée en France

[Poibeau & Nazarenko 99]

EXTRACTION D'INFORMATION
Deux phases de travaux

Contexte en 1997-98

- Des systèmes à base de règles éprouvés mais « figés »
- Un coût élevé d'adaptation à de nouveaux domaines
  - ➢ Recourir aux méthodes d'apprentissage

[Soderland 95, Riloff 96, Yangarber & Grishman 97, Freitag 98]

Premiers travaux

- Thèse de T. Poibeau (1998-2002, Cifre Thales)
- Collaboration avec C. Nédellec (LRI puis MIG-INRA)

Application à la génomique (LIPN-MIG)

- Projets « BioInformatique » Caderige 1 et 2 (IMAG, 2001-03)
- Projet RNTL ExtraPloDocs (LIPN, 2002-05)
- Projet européen STREP Alvis (HUT, Finlande, 2004-2006)

[Bessières *et al*. 01, Nédellec & Nazarenko 01, 04, Alphonse *et al*. 04]

REPÉRER DES INTERACTIONS ENTRE GÈNES DANS LES TEXTES DE GÉNOMIQUE

[Collier *et al*. 99, Blaschke *et al*. 99, Rindflesh et al.00 , Ono 01, Marcotte *et al*. 01, Thomas *et al*. 01, Nédellec & Nazarenko 01]

Fragment d'un résumé de MedLine

`[..] the GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK [..]`

Formulaire

Interaction	**Type : negative**		
	Agent : GerE protein		
	Cible :	**Expression :**	**Source : sigK gene** **Produit : sigmaK**

EXPLOITER DES INDICES LINGUISTIQUES

GerE stimulates cotD transcription and y cotA transcription […], and, unexpectedly, inhibits […] transcription of the gene (sigK) […]

Les règles reposant sur des indices de surface produisent du bruit

<nom de gène/protéine> inhibits <nom de gène/protéine>

~~**cotD inhibits sigK**~~

Il faut exploiter des règles comportant des contraintes linguistiques

SI	le sujet X d'un **verbe Y** d'interaction est un nom de protéine et l'objet direct Z est un nom de gène ou l'expression d'un gène	**ALORS**	Il y a une **interaction** dont **X est l'agent** et **Z est la cible**

COMMENT ÉCRIRE CES RÈGLES ?

Les difficultés classiques

- Les règles écrites manuellement sont insuffisantes
  - ✓ Les règles ont une bonne précision (> 80 %)
  - ✖ Les formes d'expression sont variables
  - ✖ Les règles ont un faible rappel
- Il faut de nouvelles règles pour toute nouvelle application

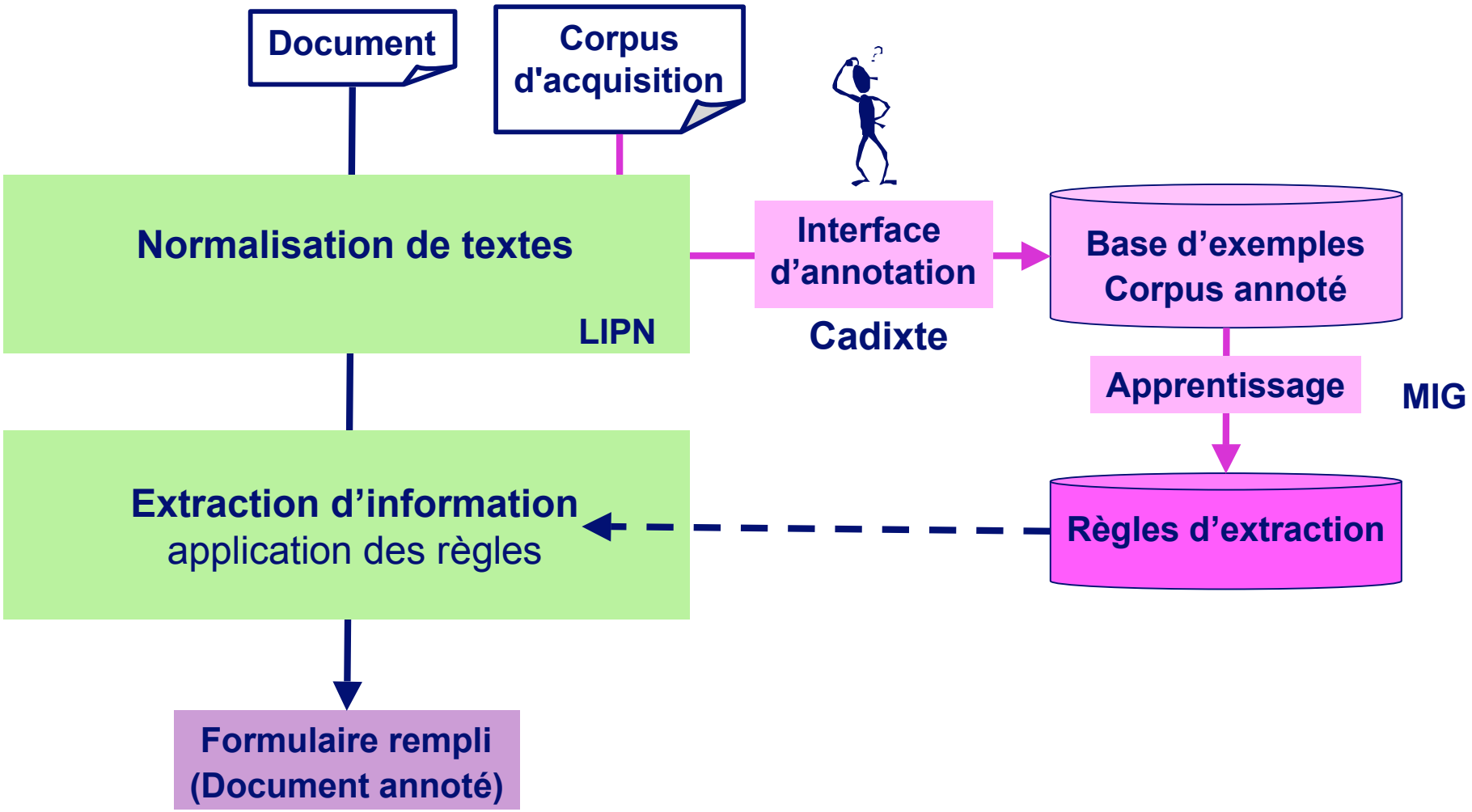
Des difficultés supplémentaires liées à l'application

- Analyser des phrases complexes
- Extraire des relations (interactions entre gènes et protéines)

L'approche : combiner apprentissage et TAL [Nédellec & Nazarenko 01]

- ➢ Apprendre les règles d'extraction
- ➢ S'affranchir au préalable de la diversité des formes de surface par une normalisation préalable des textes

NORMALISATION DE TEXTES ET APPRENTISSAGE



LES DIFFICULTÉS DE L' ANALYSE SÉMANTIQUE

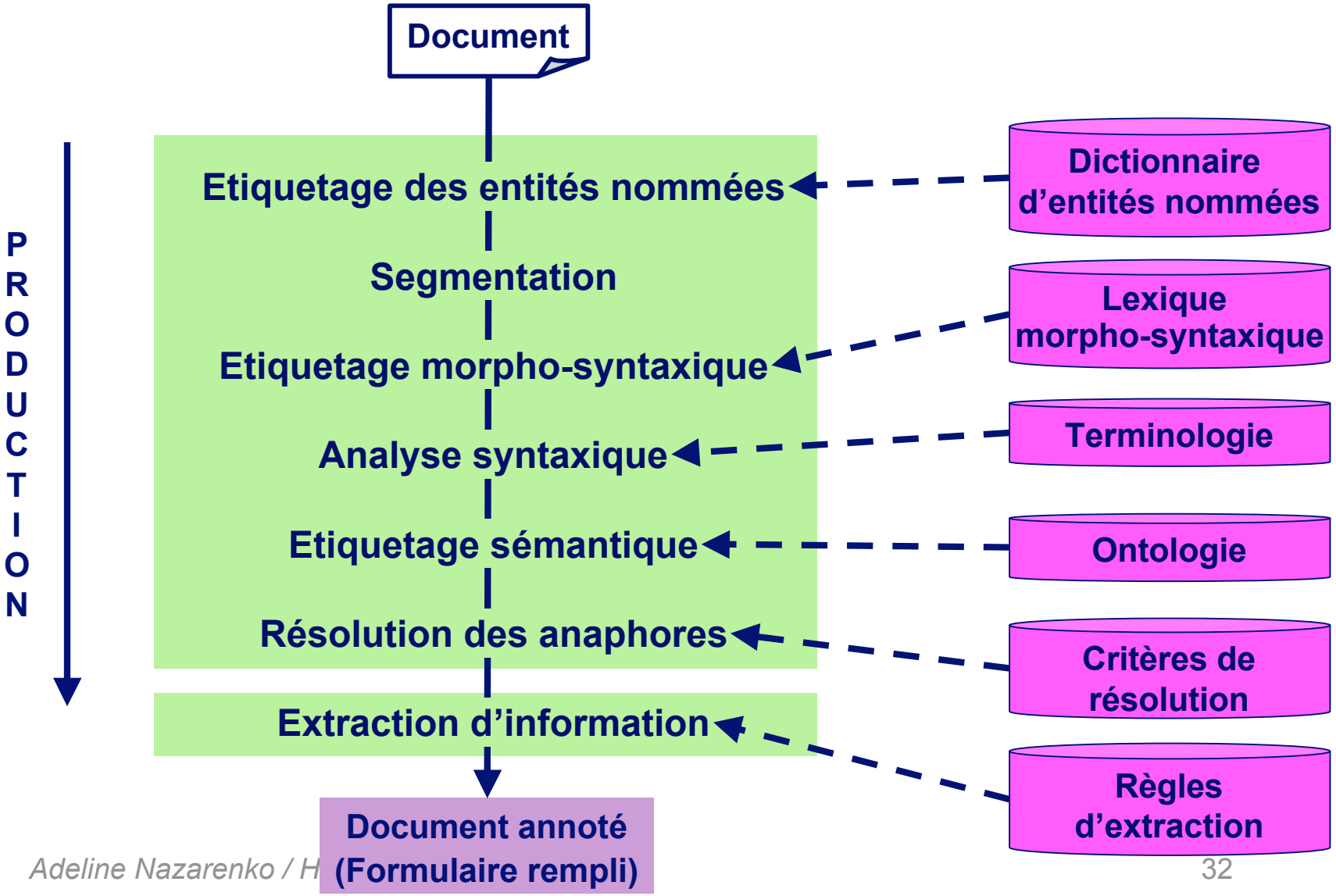
Réduire la diversité des formes de surface

- **Les normaliser suppose différents niveaux d'analyse**
  - Etiquetage morpho-syntaxique
  - Analyse syntaxique : syntagmes, dépendances
  - Analyse référentielle : référents du discours, coréférence
  - Etiquetage sémantique : catégories sémantiques
  - Mesure de saillance : différents facteurs

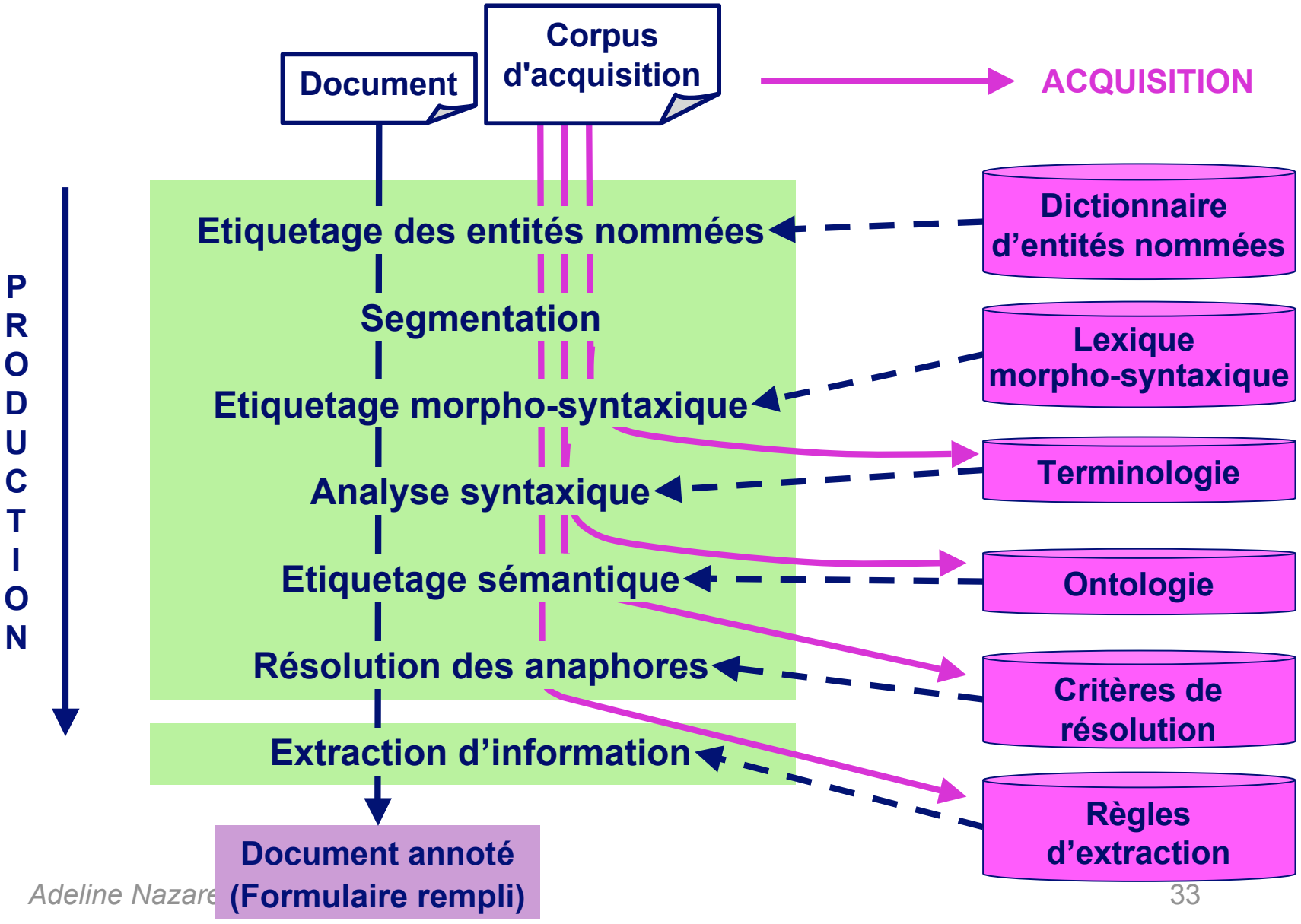
Tenir compte des effets contextuels

- **Adapter l'analyse en fonction des 3 niveaux de contexte**
  - Le domaine de spécialité → acquisition à partir de corpus
  - Le texte dans son ensemble → apprentissage endogène
  - **Les modalités (non prises en compte)**

NORMALISER LES TEXTES



COUPLER ACQUISITION ET PRODUCTION



ENJEUX DU TRAVAIL EN COURS

[Alphonse et al. 04]

- **Valider l'architecture et le couplage Production - Acquisition** (MIG-LIPN)

- **Mettre au point des outils d'acquisition**
  - Acquisition de dictionnaires d'entités nommées [Fukuda et al. 98] [Proux *et al*. 98] [Kim *et al*. 04] (T. Poibeau)
  - Acquisition de termes (S. Aubin, T. Hamon)
  - Acquisition d'ontologie (MIG)
  - Pondération des facteurs de résolution d'anaphore (D. Weissenbacher)
  - Apprentissage de règles d'extraction (MIG)

- **Développer des nouveaux outils d'analyse :** étiqueteur, analyseur...

- **Eprouver l'adaptation de l'analyse au domaine d'application**
  - **Analyse syntaxique (S. Aubin)**
  - **Résolution d'anaphores (D. Weissenbacher)**

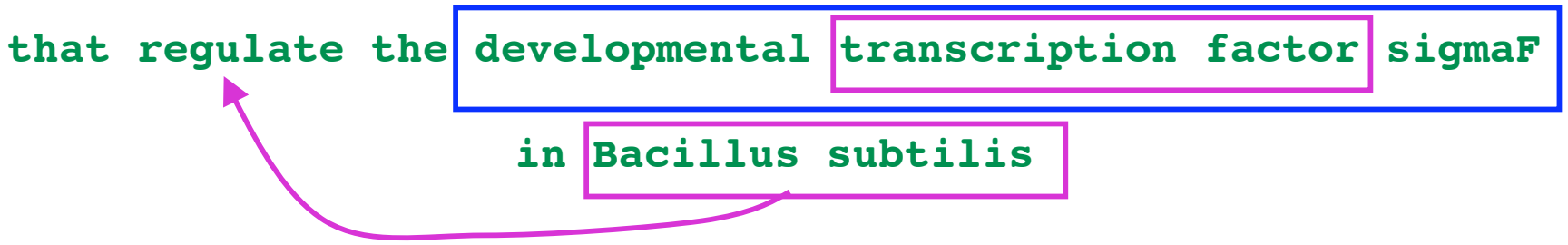
ADAPTATION DE L'ANALYSE SYNTAXIQUE

Bilan des analyseurs génériques (IFSP, Link parser)

- Faible qualité sur des corpus complexes très spécialisés
- Exemple : 25 % d'erreurs pour les relations sujet-verbe ! [Aubin 03]

Difficultés

- Groupes nominaux complexes
- Ambiguïtés de rattachement



Intégration de connaissances terminologiques

Gain significatif pour certaines relations : - 40 % [Cohen 01]

POURQUOI RÉSOUDRE LES ANAPHORES ?

The operon of Bacillus subtilis consists of the genes hrcA, grpE [...]. It is controlled by the CIRCE/HrcA operator/repressor system...

Objectif de l'extraction d'information

- Identifier les référents du discours
- Les mettre en relation

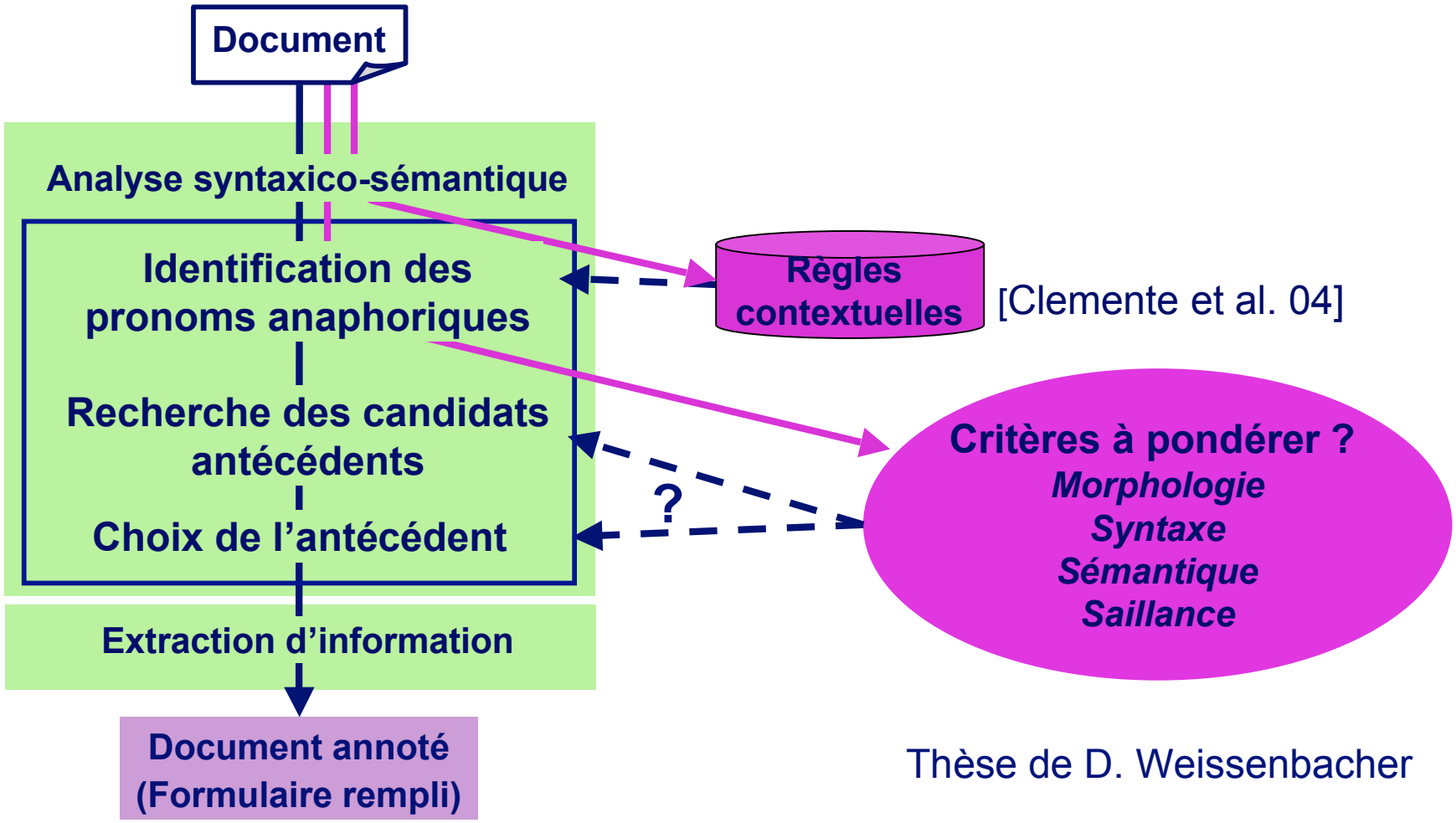
Quels référents ?

- On se focalise généralement sur les entités nommées : **hrcA**
  - Eléments faciles à repérer
  - Eléments saillants à identifier
- Mais ce ne sont pas les seuls éléments intéressants
  - Pronoms anaphoriques : **It** (dans 2/3 des phrases [El Zant 02])
  - Expressions définies : **The operon of Bacillus subtilis**

Résoudre les anaphores permet de

- Enrichir la description des référents du discours
- Augmenter le rappel de l'extraction d'information
- Mieux tenir compte des fréquences dans les calculs de pertinence

RÉSOLUTION DES ANAPHORES
Nécessité d'une adaptation ?



Thèse de D. Weissenbacher

CONCLUSION

Il faut des connaissances pour analyser les textes

C'est d'abord à partir des textes qu'on acquiert des connaissances

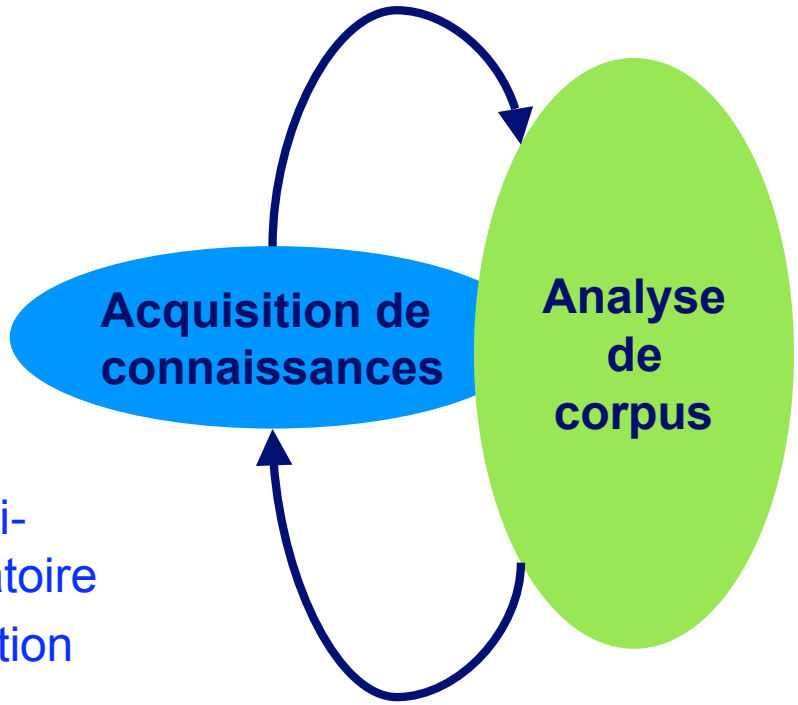
Diverses expériences d'acquisition

Une démarche d'acquisition

Défis

Techniques d'expéri-mentation en laboratoire

Des outils de validation des résultats



Une analyse paramétrable selon l'application

Collaboration entre TAL et apprentissage

Aller plus loin dans la normalisation

Analyse référentielle

Modalités

PERSPECTIVES

Concevoir des outils d'exploration de documents

- Indexation de sites web
- Hypertextualisation, outils de navigation
- Indexation spécialisée des documents médicaux

Explorer une nouvelle piste : l'ancrage linguistique d'ontologie pour la fusion d'ontologies

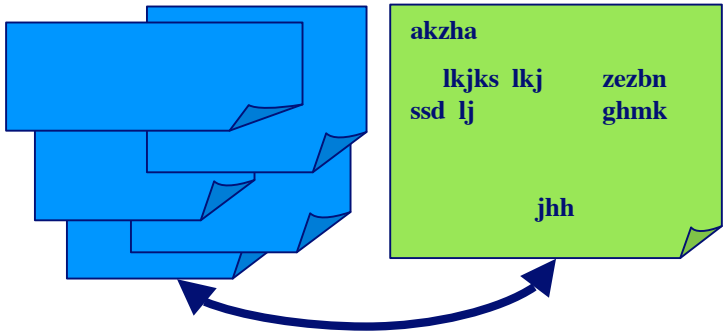
Articuler deux sémantiques différentes

Acquisition de connaissances

Analyse de corpus

Analyse globale sur le texte ou le corpus

Sémantique distributionnelle et textuelle



Analyse locale de fragments de texte

Sémantique référentielle (entités nommées)