

Du texte brut au web sémantique

Thierry Poibeau

LIPN, CNRS et Université Paris 13



Faciliter l'accès au texte

- But : Améliorer l'accès à l'information (essentiellement textuelle)
- Moyen : Proposer de nouveaux outils d'accès au texte
 - Annotation du texte
 - Index structurés, ontologies
 - Modèles formels du texte
- Pré-requis : normaliser le texte, passer du texte à un format structuré

Plan de la présentation

- Quelques applications
- Techniques de reconnaissance de séquences linguistiques
- Techniques de mise en correspondance de séquences
- Conclusion

Exemples d'application

Annotation de documents

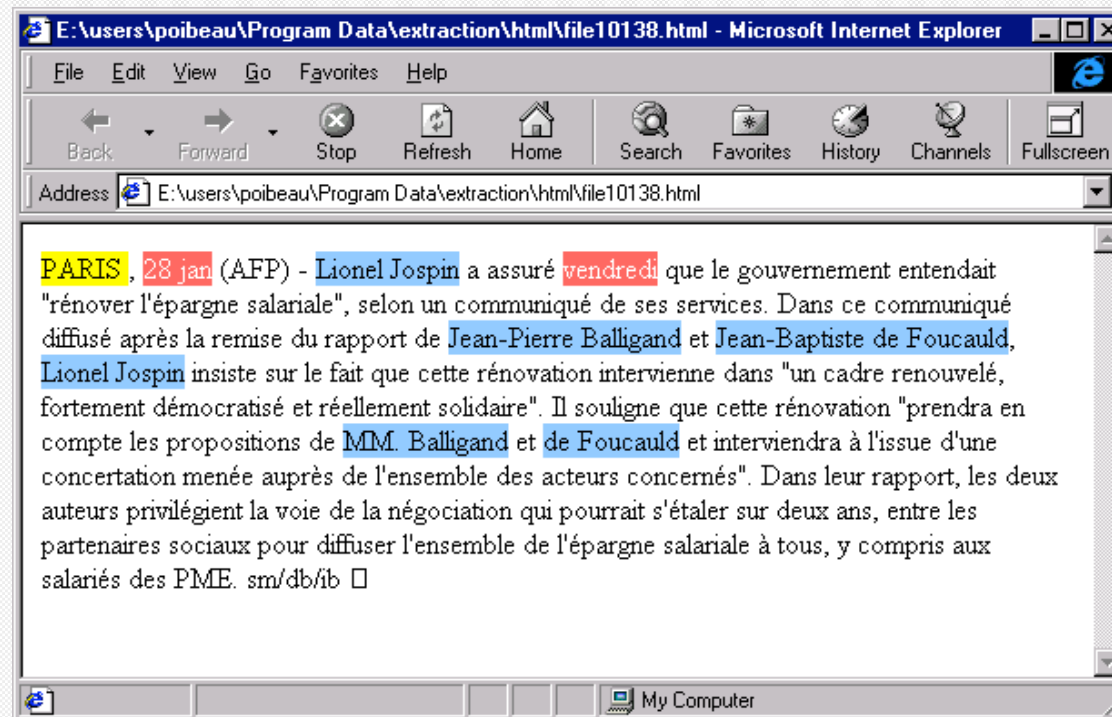


Plate-forme d'annotation sémantique

Accès à des bases multilingues

The image displays three overlapping screenshots of Microsoft Internet Explorer, each showing a different language interface with named entities highlighted in red. The top-left window shows Arabic text with entities like "الكاميرون" and "بهارين". The middle window shows Russian text with entities like "США", "президент", and "Ирак". The bottom-right window shows Polish text with entities like "Rada Bezpieczeństwa ONZ", "Iraku", "Szwedzki dyplomata Hans Blix", and "Bagdad". Each window also shows the browser's address bar, navigation buttons, and a search engine (Google).

Entités nommées sur textes multilingues (INaLCO 2003)

Extraction d'information

(Remplissage d'une base de données)

The screenshot shows a Windows desktop environment. In the background, a Notepad window titled 'afp_0001.txt - Bloc-notes' contains a news article snippet: 'PARIS, 5 oct (AFRP) - Le Premier ministre français Lionel Jospin a affirmé jeudi que "M. I éviter u peuple".' Below this, another snippet reads: 'Vojislav l'Opposi (DOS), s libérée" manifest à Belgra fier d'a Yougosla centaine dans les terrasse'.

In the foreground, two data extraction forms are visible. The first form, titled 'e:\users\poibeau\program data\extraction\txt\file10138.txt', has the following fields filled:

Date	28-ja-00
Location	PARIS
Personality	Lionel Jospin
Type	Déclaration
Source	AFP (wire)
Topic	- Lionel Josp gouvernement l'épargne sal communiqué de

The second form, titled 'e:\users\poibeau\program data\extraction\txt\file10139.txt', has the following fields filled:

Date	31-ju-00
Location	AUXERRE
Sport	football
Type	Résultat sportif
Source	AFP (wire)
Topic	- "Maintenant, il me faudra réapprendre à perdre". Lors de sa nomination pour succéder à Guy Roux comme entraîneur de l'AJ Auxerre, Daniel Rolland avait eu ce bon mot. Samedi soir à domicile, pour son

Plate-forme d'extraction d'information

Systemes de question-réponse

The screenshot shows a Mozilla Firefox browser window displaying a web application titled "Système de Questions-Réponses: DESS IM 2005". The interface includes a menu bar with "Terminologie", "Prédicats", "Questions", and "Questions-Réponses". Below the menu, there are tabs for "Questions-Réponses" and "Corpus / Scripts". The main content area is divided into two panes: "Fichier Fichiers/reco" and "Système de Questions-Réponses: DESS IM 2005". The left pane shows XML code, and the right pane shows a document titled "Schizophrénies débutantes : diagnostic et modalités de prise en charge". A smaller window in the foreground displays a question and its answer.

XML code (left pane):

```
- <INDEXGLOBAL>
- <INDEXPRED>
- <EXPRESSION>
  <PREDICAT>datation</PREDICAT>
  <ARGUMENT1>L'apparition de la maladie</ARGUMENT1>
  <ARGUMENT2>avant 25 ans</ARGUMENT2>
</EXPRESSION>
- <EXPRESSION>
  <PREDICAT>def</PREDICAT>
  <ARGUMENT1>L'information</ARGUMENT1>
  <ARGUMENT2>un outil thérapeutique in</ARGUMENT2>
  favorisant</ARGUMENT2>
</EXPRESSION>
- <EXPRESSION>
  <PREDICAT>def</PREDICAT>
  <ARGUMENT1>DSM IV</ARGUMENT1>
  <ARGUMENT2>la classification américai</ARGUMENT2>
</EXPRESSION>
- <EXPRESSION>
  <PREDICAT>def</PREDICAT>
  <ARGUMENT1>CIM 10</ARGUMENT1>
  <ARGUMENT2>La classification interna</ARGUMENT2>
</EXPRESSION>
- <EXPRESSION>
  <PREDICAT>def</PREDICAT>
  <ARGUMENT1>DUP</ARGUMENT1>
  <ARGUMENT2>la durée de psychose non</ARGUMENT2>
  traitée</ARGUMENT2>
</EXPRESSION>
- <EXPRESSION>
  <PREDICAT>def</PREDICAT>
```

Document content (right pane):

Schizophrénies débutantes : diagnostic et modalités de prise en charge
Conférence de consensus
organisée par
Fédération Française de Psychiatrie
selon la méthodologie de IANAES
avec le soutien de la Direction Générale de la Santé
23 et 24 janvier 2003

Membres du Jury : Dr François PETITJEAN (Président), Dr Marie-CARDINE (Présidente du Jury), Dr M. BA...
Melle C. CABANAC, Mme C. CHERIAUX-FALLOU...
SSE, Mme De MAXIMY, M B. ESCAIG, Dr B. EST...
GURY, Dr C. LAUNAY, Mr Ch. MARCHANDET, M...

Texte des recommandations longues élaboré par le jury
Introduction

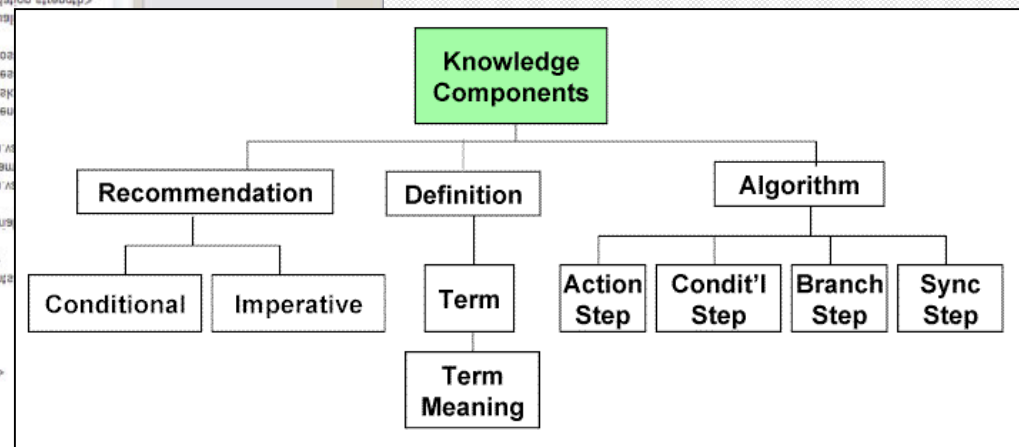
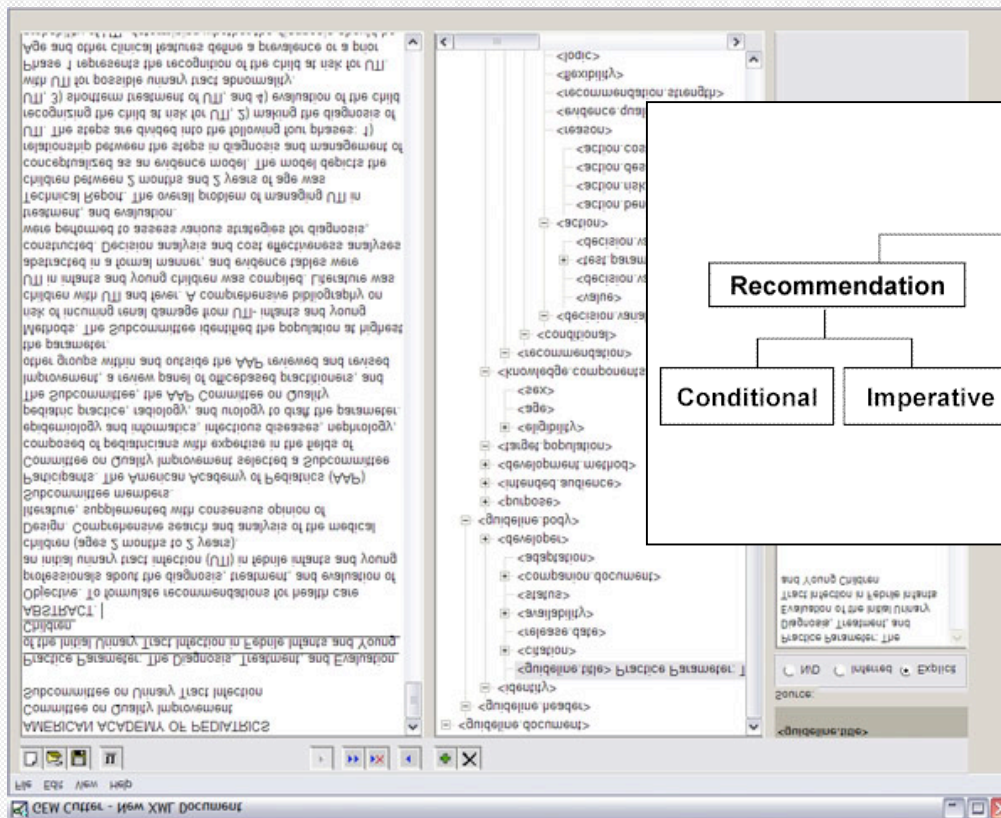
Question (foreground window):

qu'est-ce que la DUP ?
quelle est la signification de DUP ?
quelle est l'équivalence de DUP ?
comment se caractérisent les psychoses délirantes aiguës ?
par quoi se caractérisent les psychoses délirantes aiguës ?
qu'est-ce qui caractérise les psychoses délirantes aiguës ?
quelle est la caractéristique des psychoses délirantes aiguës ?
par quoi les psychoses délirantes aiguës sont-elles caractérisées ?

Réponse(s):
la durée de psychose non traitée

Systeme de QR médical – INaLCO 2005

Aide à la modélisation



Modélisation des guides de bonne pratique médicale – DTD GEM (Yale)

Travaux du groupe ActMed (PPF P13) et de A. Bouffier (LIPN)

Niveaux d'analyse

- Analyse de séquences isolées
 - Signifiante du signe (sémiologie)
- Extraction d'information
 - Signifiante du discours (sémantique)
- Analyse textuelle
 - Signifiante du texte (métasémantique)
- Cf. Au-delà du web sémantique, des travaux de linguistique traditionnelle (Saussure, Bakhtine, ...)
- Voir aussi Ogden&Richards

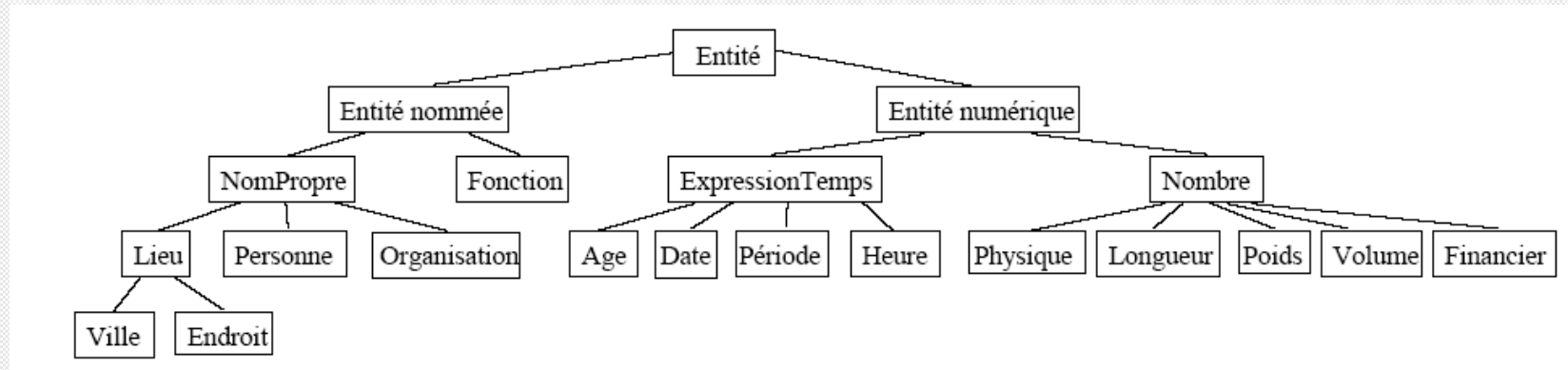
Analyse de séquences isolées

Base de l'analyse

- Indexation de séquences pertinentes
 - Terminologie
 - Entités nommées
 - Classes sémantiques
- Besoins de ressources spécialisées
 - Lexiques existants
 - Stratégies d'adaptation à la tâche

Analyse des entités

- Reconnaissance les séquences pertinentes
- Typage par rapport à une ontologie
- Normalisation



Stratégies

- Plusieurs approches
 - Dictionnaires et ensemble de règles
 - Apprentissage à partir de données annotées
 - Approches mixtes
- Quelle approche obtient les meilleurs résultats ?
 - Résultats comparables (cf. IREX, ...)
 - Conditions de mise en œuvre (disponibilité de données annotées ou non, type de texte analysé, etc.)

Analyse par règles

■ Étiquetage lexical

Reconnaissance des nombres,

Reconnaissance des noms propres (listes de prénoms, de lieux...)

Reconnaissance des amorces *M.* (*Monsieur*) ou *SA* (*Société anonyme*)

Reconnaissance et normalisation des sigles comme *I.B.M*

Analyse des mots inconnus et des mots commençant par une majuscule...

■ Règles de regroupement

$\$Clé_hydro \$Article_min+ [\$NP+]Hydronyme$

rives de la Kamogawa

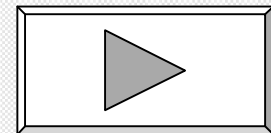
$\$Titre_militaire \$Adj_nationalité?[\$Prénom*\$NP] Patronyme$

commandant Massoud

(Fourour 2002, cf. aussi Trouilleux 2002, Poibeau 2003)

Systemes hybrides

- L'étiquetage par un système à base de règles produit des « connaissances »
- Analyse de ces connaissances par apprentissage (statistique ou symbolique)
 - Repérage de structures de discours
 - Correction dynamique d'étiquettes par défaut
- Phase de « réétiquetage »



Évaluation



- Comparaison avec une « référence »
 - Corpus journalistiques (MUC, ACE...)
 - Corpus variés (transcription de l'oral...)
- Participation à la campagne ESTER
 - Transcription de journaux radiodiffusés 3 systèmes ont participé
 - LIPN : P&R $\approx 0,65$ (0,9 P&R sur MUC)
(P&R = moyenne harmonique précision et rappel)
 - Complémentarité des méthodes (règles + apprentissage)

Annotation sémantique

Étiquetage sémantique

- Enrichissement de documents par analyse sémantique
 - Mise en évidence de mots clés
 - Catégorisation, classification de textes
- Nécessité de gérer la variation sémantique (synonymes)
- Deux types d'approches
 - À partir d'une base de connaissance pré-définie
 - À partir d'une phase d'apprentissage sur corpus

Comparaison des deux approches

- Expérience sur un corpus financier
- Deux approches sont comparées
 - Acquisition sur corpus à partir du logiciel Asium (Faure 2000, analyse distrib. en corpus) 
 - Adaptation des ressources issues du réseau sémantique Lexidiom (Memodata) 
- Complémentarité de ces ressources
 - Mots essentiels manquant au Lexidiom
 - Analyse à partir de corpus trop bruitées

Évaluation d'une approche hybride

- Combiner les approches
 - Ressources externes : guider et filtrer l'analyse à partir de corpus
 - Apprentissage automatique : capturer les termes spécifiques au domaine
- Résultats
 - Gain de temps : convergence beaucoup plus rapide vers les éléments pertinents du corpus
 - Gain de performance : meilleur rapport rappel/précision que chaque ressource isolée

Mise en relation d'entités

Principe de l'extraction

- Produire une forme normalisée du texte, indépendamment de la variation linguistique

11983. LVMH achète Victoire Multimédia.

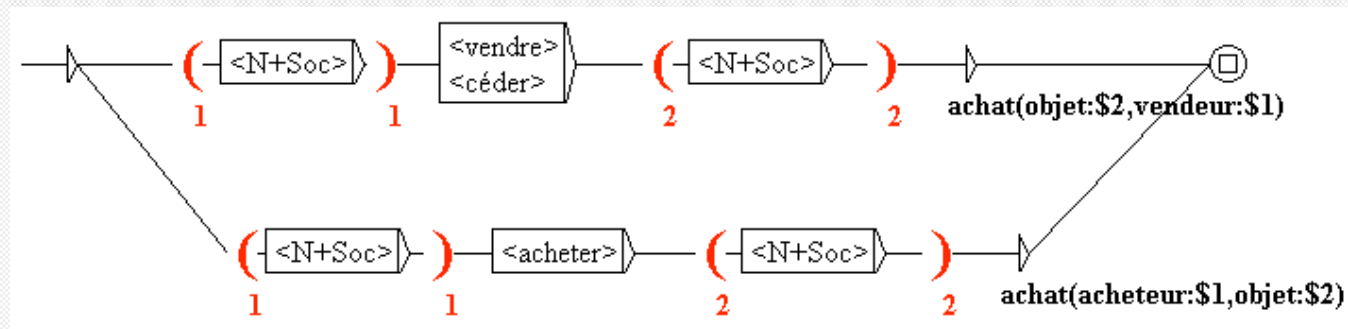
12986. Victoire Multimédia a été acheté par LVMH.

11983. LVMH, qui a récemment acheté Victoire Multimédia...

achat(objet:Victoire Multimédia, acheteur:LVMH)

Approche classique

■ Définition manuelle de transducteurs



■ Limites

- Caractère *ad hoc* des ressources
- Peu réutilisable, peu fiable (problèmes de couverture)

Automatiser l'approche

- Apprentissage semi-automatique de règles d'extraction
- Nécessité de multiples outils
 - Reconnaisseurs d'entités, de termes, ...
 - Analyseurs syntaxiques, sémantiques, ...
- Nécessité de ressources importantes
 - Dictionnaires et lexiques spécialisés ou non
 - Outils d'acquisition et d'adaptation

Travaux menés au LIPN

- Annotation linguistique multi-niveaux de corpus
 - Plate-forme d'annotation
 - Projet ALVIS (A. Nazarenko, T. Hamon, S. Aubin, J. Derivière, D. Weissenbacher...)
- Ressources syntaxico-sémantiques
 - Acquisition de schémas prédicat-arguments (F. Gayral, L. Audibert, A. Bossard, T. Poibeau)
- Apprentissage de règles d'extraction
 - Acquisition à partir des couches d'analyse linguistique (E. Alphonse + INRA/MIG)

Annotation linguistique multi-niveaux

Exemple de la biologie

(transparents élaborés en
collaboration avec A. Nazarenko)

Structure des règles apprises

- Ensemble de règles d'extraction (transducteurs)
SI <conditions> ALORS <actions>

**Si le fragment de texte considéré vérifie
l'ensemble des conditions**

**Alors les actions qui permettent de remplir
tout ou partie du formulaires sont
déclenchées**

Limites des approches superficielles

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

Présence de 2 noms de gènes/protéines + critères stat.

[Pillet 00, Nédellec *et al.* 01]

- 80% Précision/Rappel pour la sélection de fragments
- Aucune information sur la nature de l'interaction

Présence d'un verbe d'interaction entre les 2 gènes/protéines [Ono *et al.* 01]

- Plus d'information, précision faible

Nécessité de règles complexes

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

Exemple de règle [Appelt et al. 1993, Grishman 1995]

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
----	---	-------	--

Apprendre des règles d'extraction

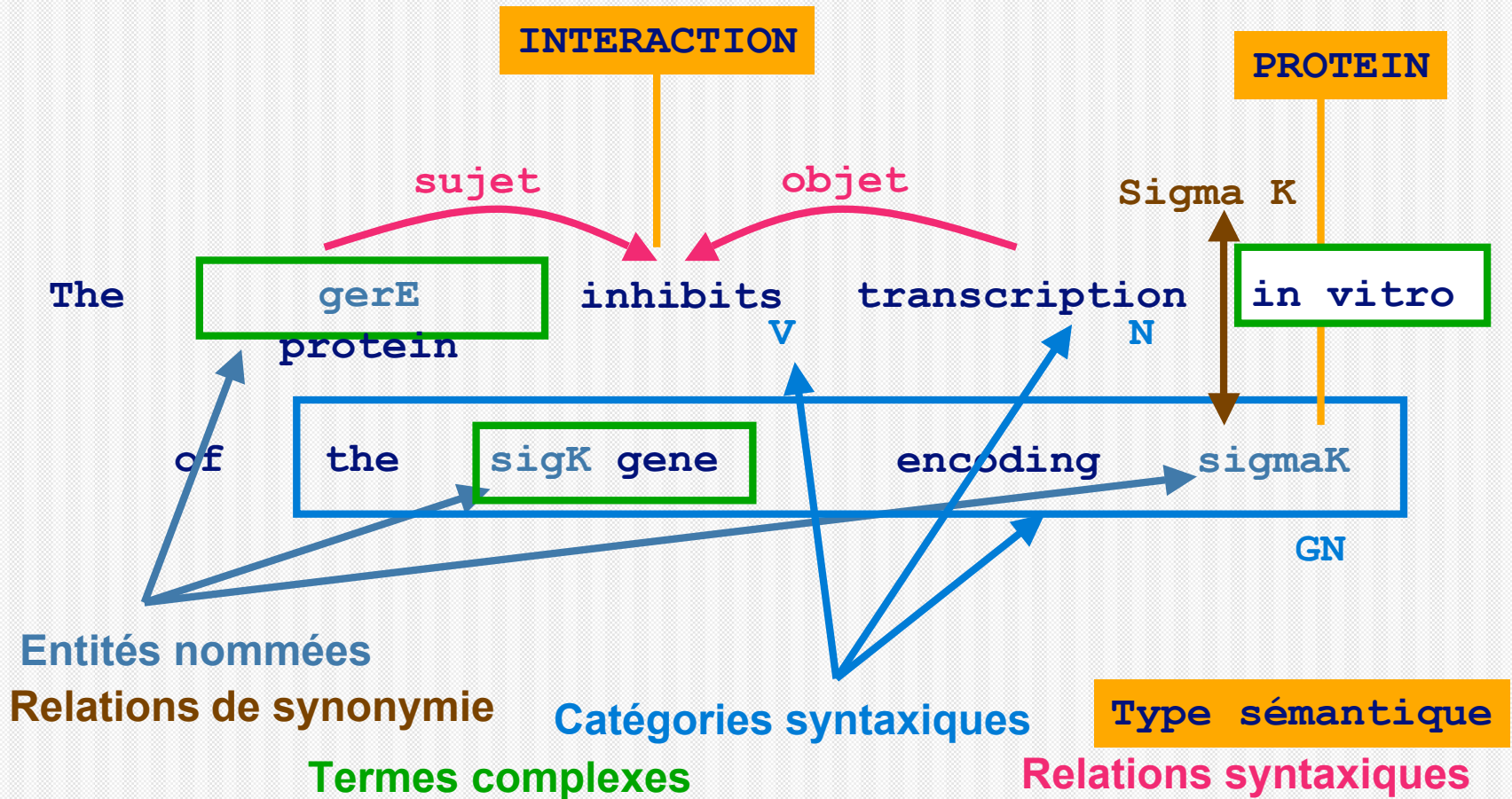
[Riloff 1996, Freitag 1998, Soderland 1995, Ciravegna 2001]

Simplifier les règles ?

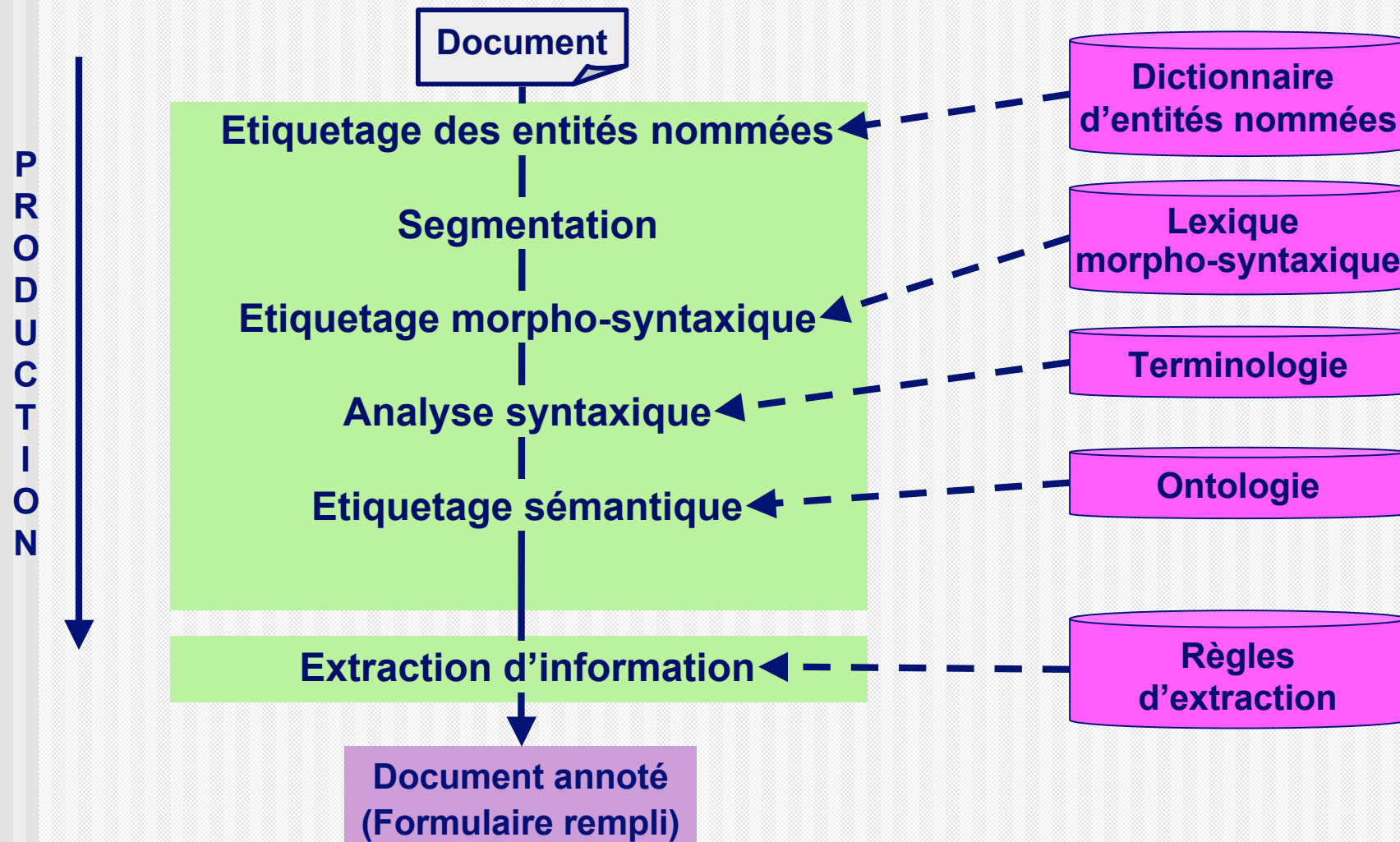
- Stratégie : supprimer une part de la variation linguistique, en amont
 - « Abstraire » la représentation (normalisation)
 - Injecter un maximum d'information sur le corpus (schémas prédicatifs, etc.)
 - Tenir compte des spécificités du corpus

Because Predicate-Argument Structures (PASs) abstract syntactical variants for the same information, patterns based on PASs are more generalized than those on surface forms of words.

Des annotations multi-couches



Architecture (annotation ling.)



Limite des outils existants

Évaluation de quelques outils génériques

Étiqueteurs	Brill, TreeTagger	« inhibits » est un verbe
Analyseurs syntaxiques	IFSP, Link Parser	« gerE protein » est le sujet de « inhibits »
Extracteurs de termes	ACABIT, SYNTEX	« in vitro » est une expression du domaine

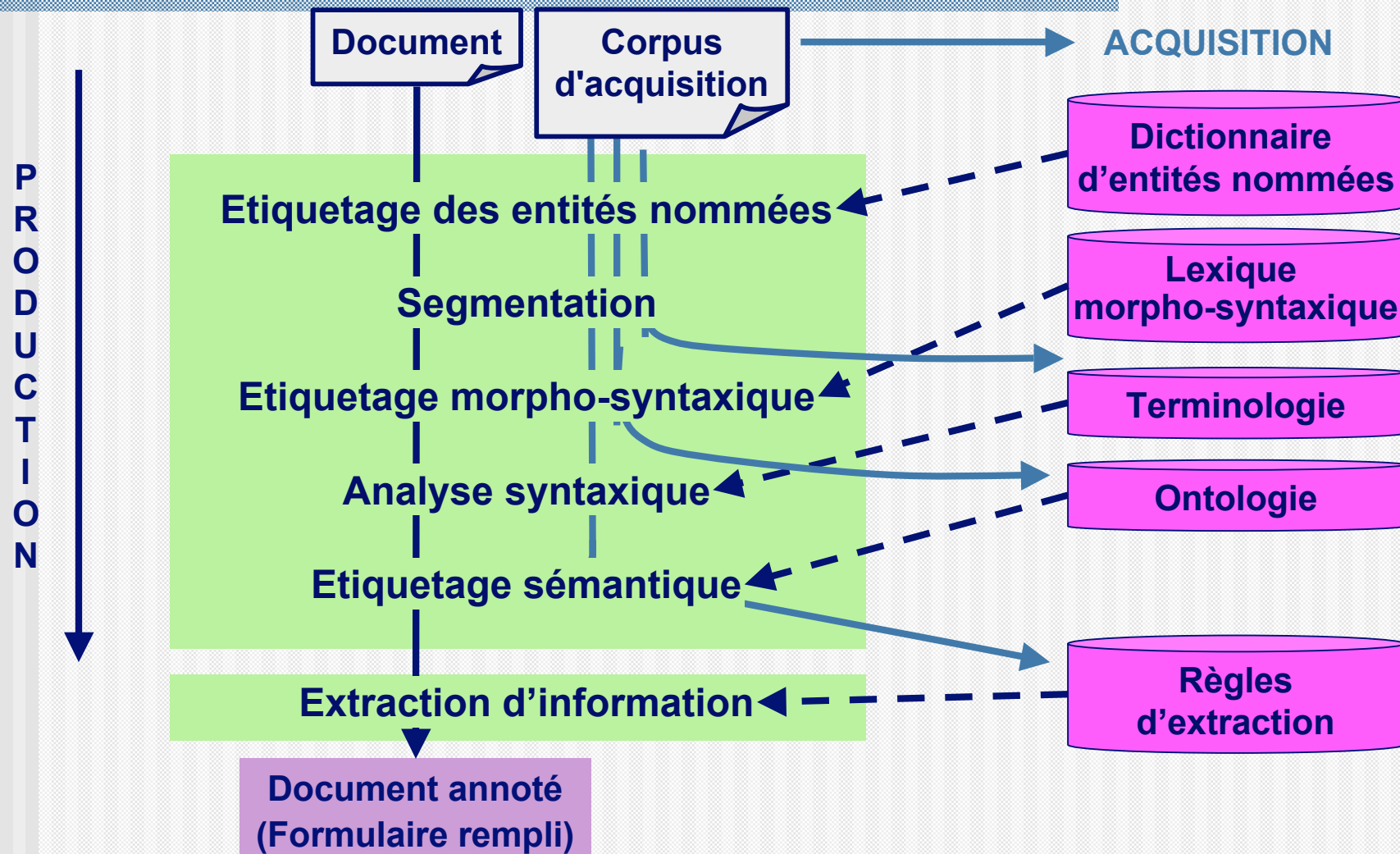
Bilan : faible qualité sur des corpus complexes aussi spécialisés

- Relation sujet-verbe : 2,5 erreurs sur 10
- Coordination multiple : 6 erreurs sur 10 !

Adaptation des traitements

- Nécessité d'adapter (rapidement) les outils au domaine et au corpus
- Stratégie
 - Exploiter le corpus pour adapter les outils
 - Injecter des connaissances : terminologie, étiquettes ou règles de grammaires spécialisées...
 - Chaque niveau utilise les informations disponibles suite aux étapes précédentes

Coupler acquisition et production



Conclusion et perspectives

Bilan

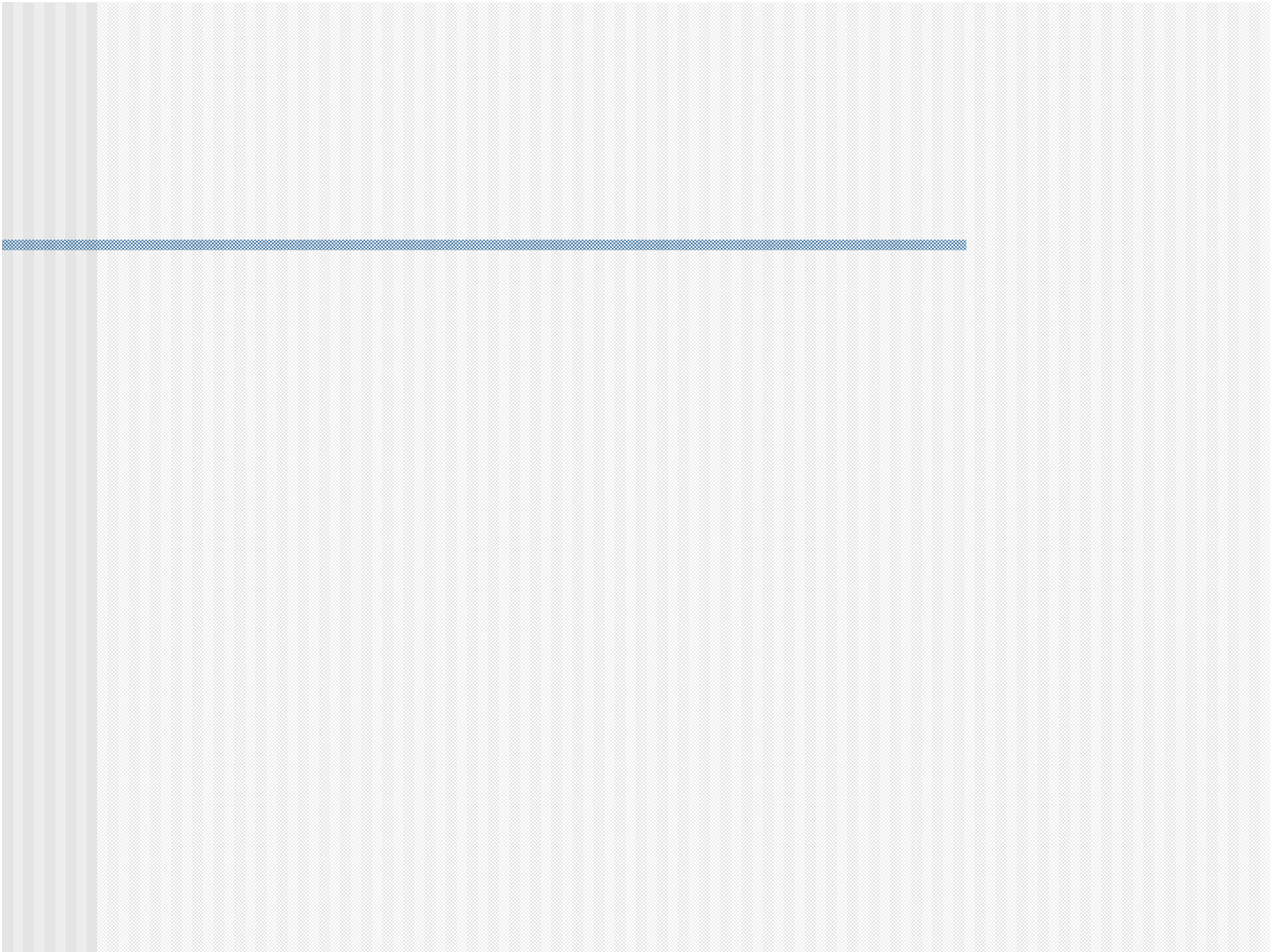
- Plate-forme d'annotation intégrant différents types d'information
 - Compatibilité partielle avec la norme ISO TC37/SC4
 - Intégration d'outils disponibles et libres, autant que faire se peut
 - Multilingue (anglais et français en cours)
- Module de traitements et stratégie d'adaptation
 - Entités nommées, extracteur de termes...
 - Adaptation du Link Parser à la biologie

Perspectives

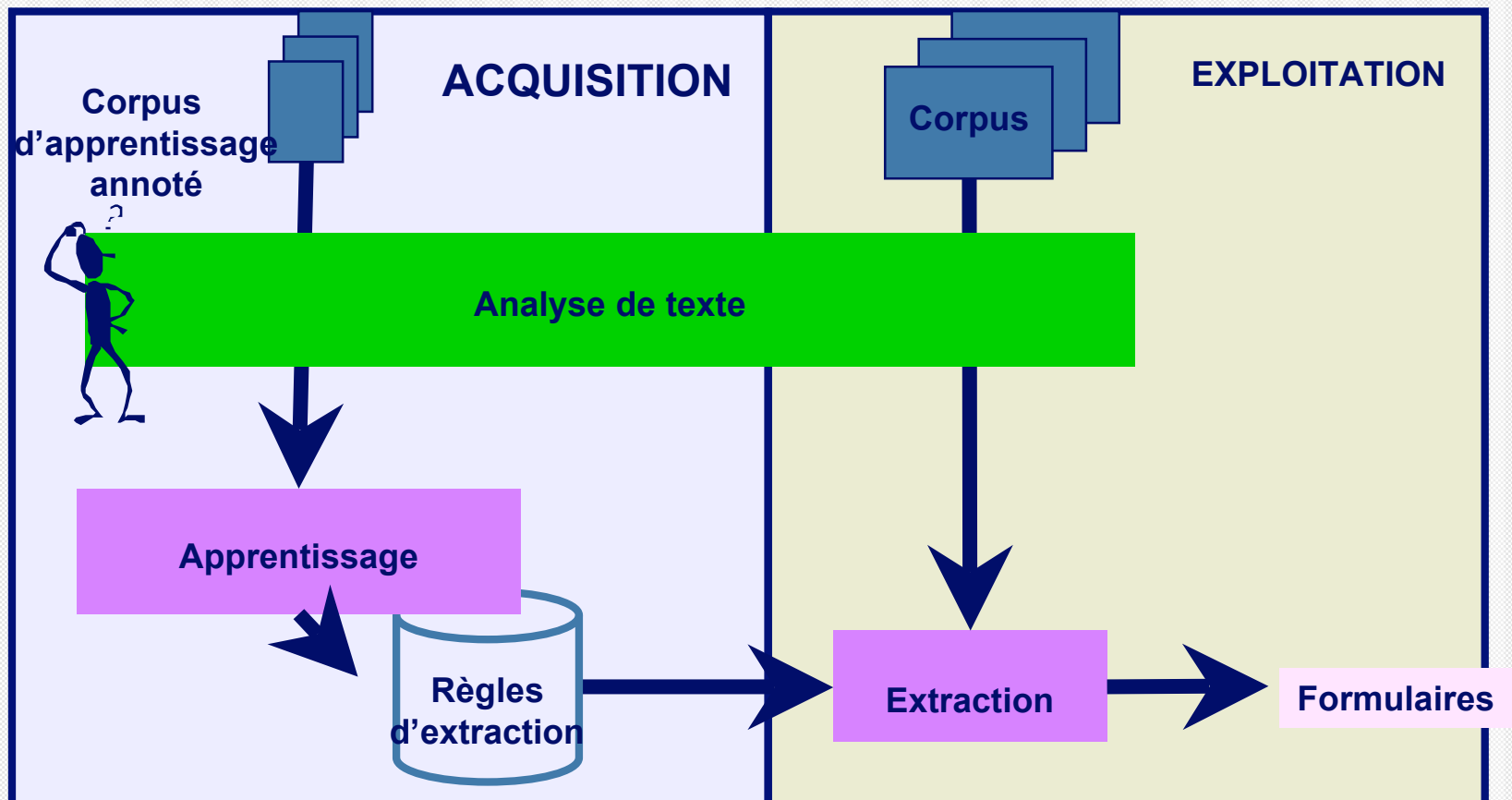
- Acquisition de schémas prédicat-arguments à base de corpus
 - Absence de ressource à large couverture pour le français
 - Lancement d'un projet de FrameNet à Nancy (Loria)
- Prise en compte du « niveau textuel »
 - Notion de typologie de textes
 - Adaptation des traitement en fonction du type de textes détecté

Fin...

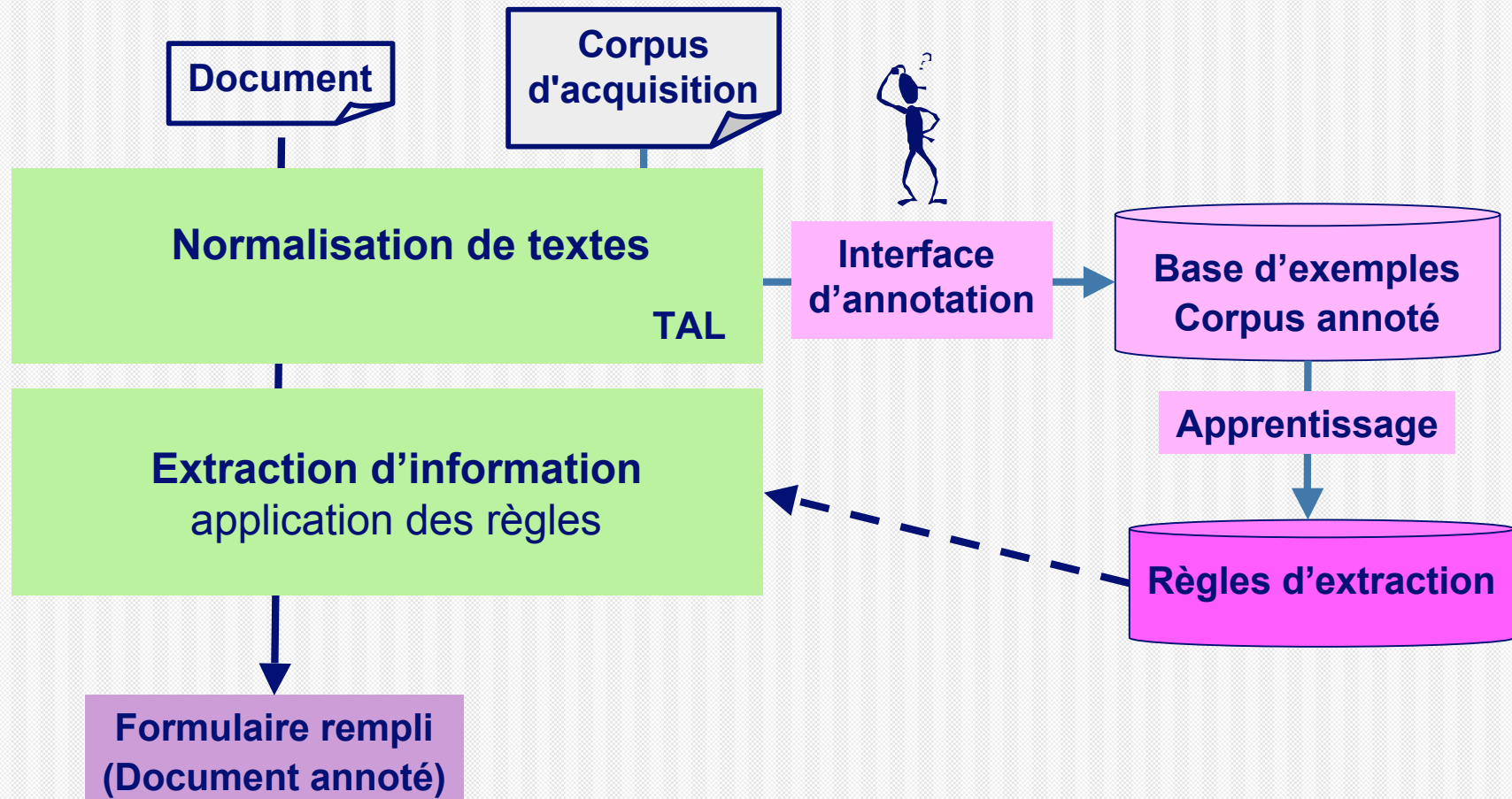
Merci aux membres de l'équipe qui travaillent sur ces projets et ces thématiques, pour leur aide et leurs conseils.



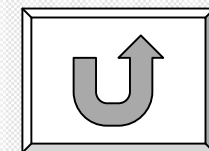
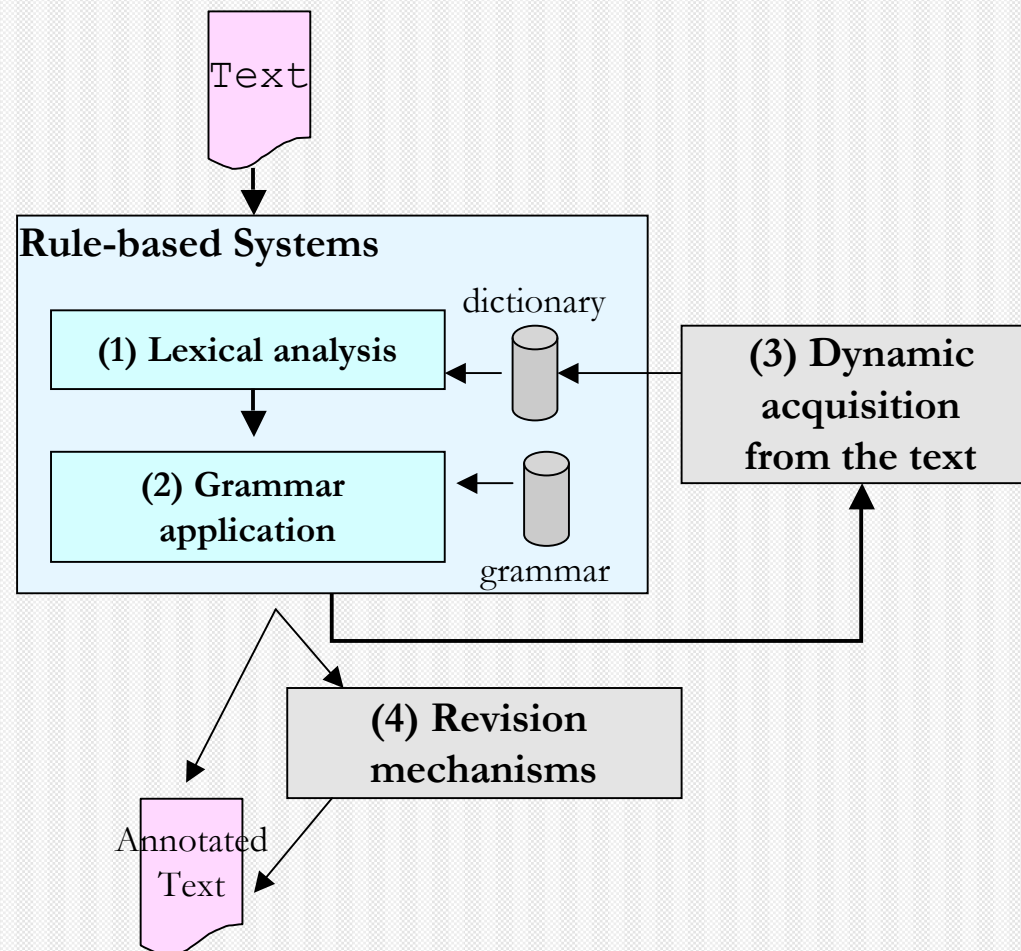
Apprentissage de règles d'extraction



Normalisation et apprentissage

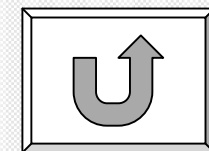
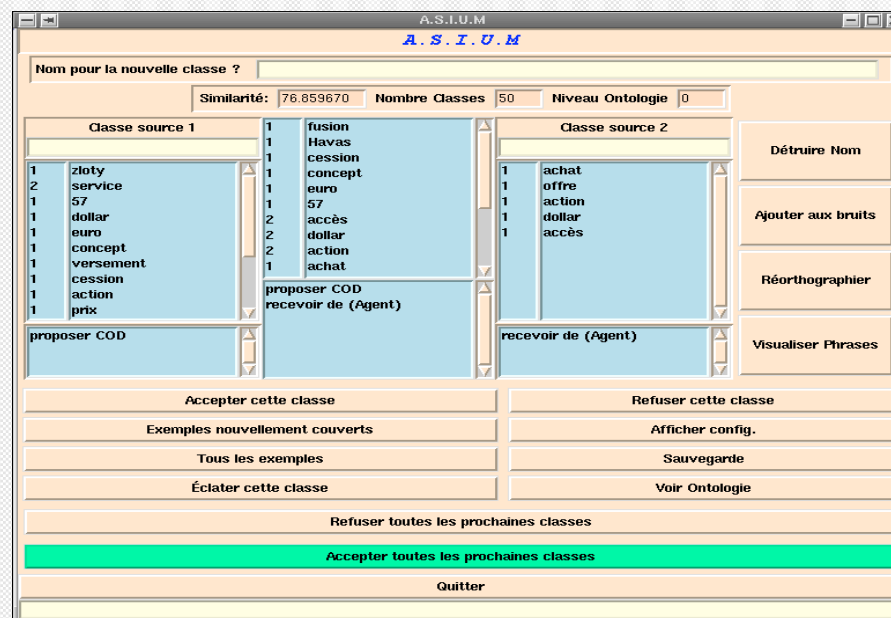


Architecture d'un système hybride



Le système Asium

- Développé par D. Faure (2000)
- Apprentissage de classes sémantiques par analyse distributionnelle de corpus



Le réseau sémantique de Memodata

- Développé par Memodata (Caen)
- Réseau sémantique du français (~ 120 000 « mots-sens »)

