# Analyse des entités nommées
# Le module du LIPN

Thierry Poibeau
LIPN – CNRS UMR 7030

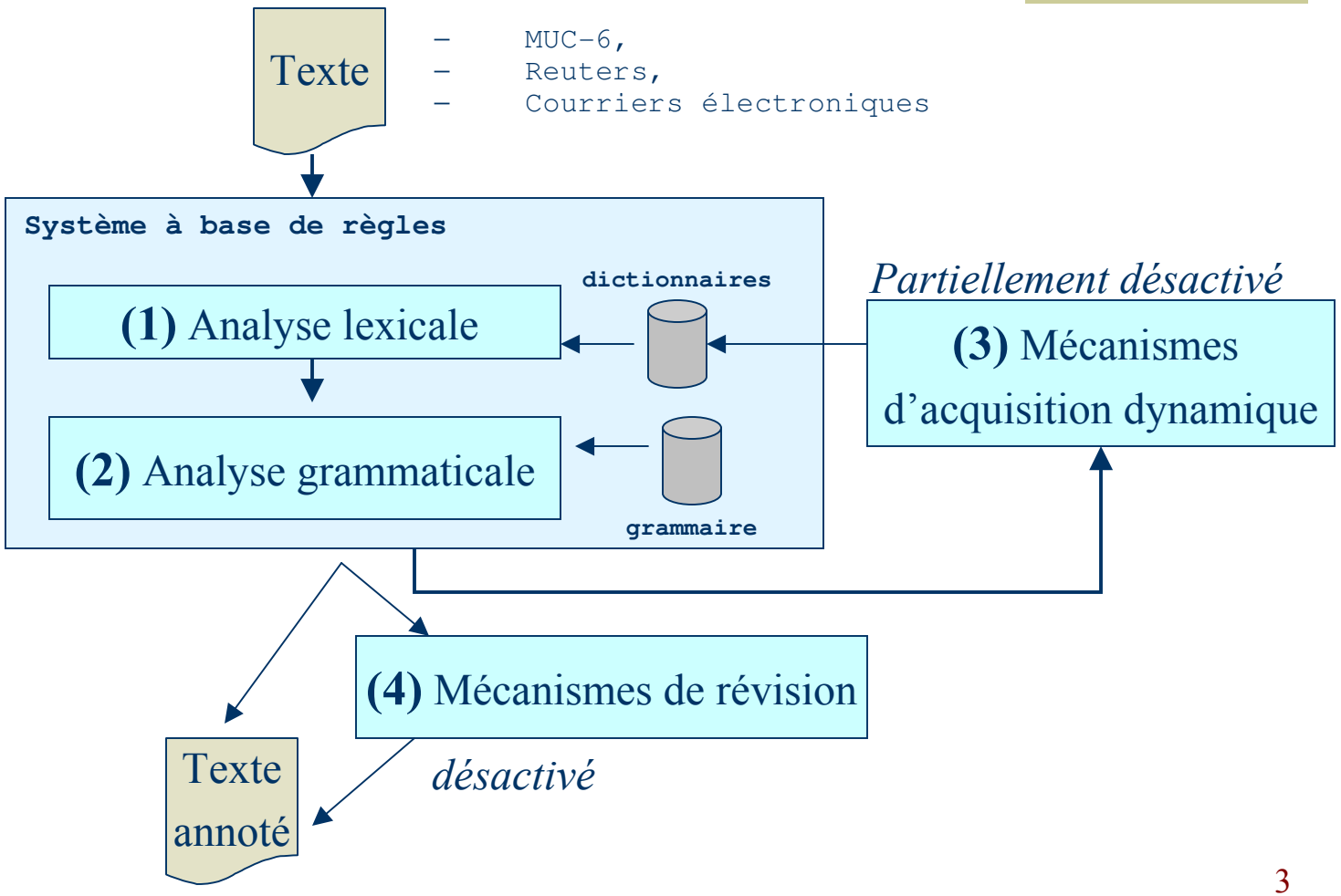
Atelier Ester 2005, Avignon le 31 mars 2005

# Le système développé

- MAEN, un système initialement développé au LIPN pour l'analyse de documents écrits
- Analyse fondé sur un ensemble de transducteurs à nombre fini d'états (cf. Unitex, Marne-la-Vallée)
- Peu d'adaptation à l'oral

➢ Évaluer les particularités du système sur des transcriptions de parole

# Architecture



# Étiquetage à partir de dictionnaires

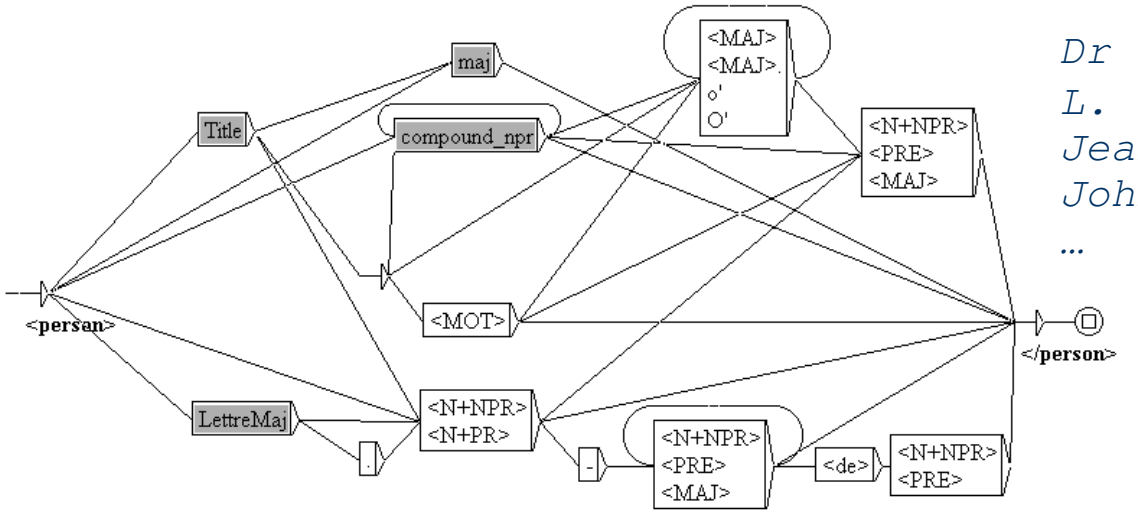
- De nombreuses sources de données, disponibles ou acquises

| Type | Étiquette | Exemple |
|---|---|---|
| Nom de personne | <N+NPR> | Dupont |
| Prénom | <N+PR> | Jacques |
| Nom de société | <N+Soc> | L'Oréal |
| Nom de géographie, dont | <N+Loc> | Picardie |
| Nom de ville[1] | <N+Loc+City> | Paris |
| Nom de pays | <N+Loc+Country> | France |

- Utilisation de valeurs par défaut (limitation de l'ambiguïté)

# Étiquetage à partir de la grammaire

- La grammaire est un ensemble de transducteurs récursifs à nombre fini d'états (>1000 états, >25000 transitions)




*Dr Jivago*
*L. Schweitzer*
*Jean-Paul Sartre*
*John M. O'Brien*
*…*

# Évaluation du système à base de règles

- Évaluation sur les transcriptions manuelles uniquement
- Performances globales médiocres (< 40 %)
  - Plusieurs catégories non couvertes initialement contribuent à faire baisser les performances (artefacts, *etc*.)
  - Distinctions difficiles (gsp/loc)
  - Adaptation insuffisante par rapport aux spécification (cf. *en 1999*)

# Conclusions et perspectives

- Campagne d'évaluation intéressante
  - Spécificités de l'oral par rapport à l'écrit
  - Qualité des ressources fournies
- Perspectives
  - Améliorer la couverture du système pour certaines catégories (*artefacts*)
  - Évaluer le système sur des transcriptions automatiques