# La chaîne d'analyse syntaxique de LEOPAR

Guy Perrier[1]

Guy.Perrier@loria.fr

Bruno Guillaume[2]

Bruno.Guillaume@loria.fr

Jonathan Marchand[1]

Jonathan.Marchand@loria.fr

[1]Nancy Université

[2]INRIA Nancy Grand-Est

Journée de l'ATALA : Quels analyseurs syntaxiques pour le français ?

# Sommaire



# Plan

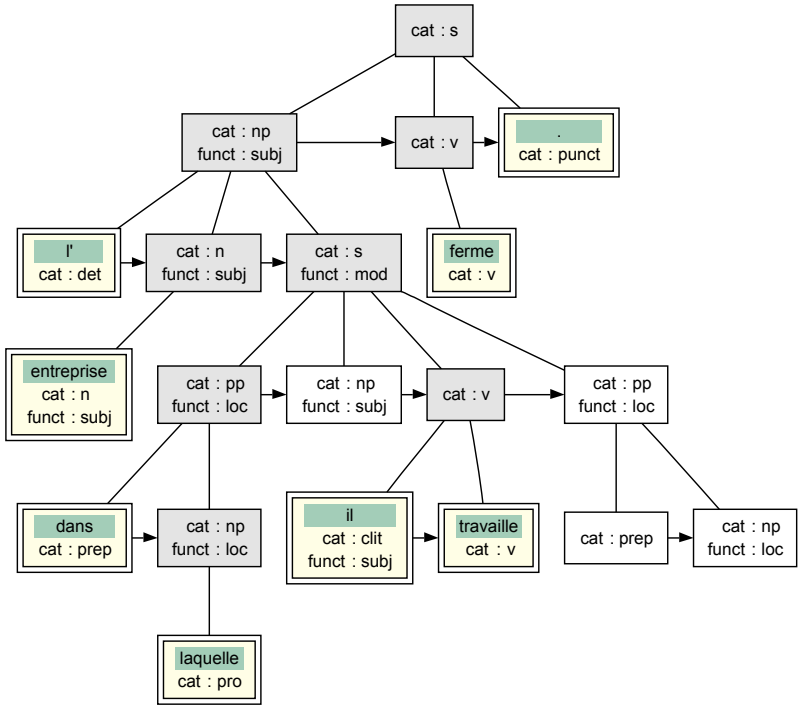


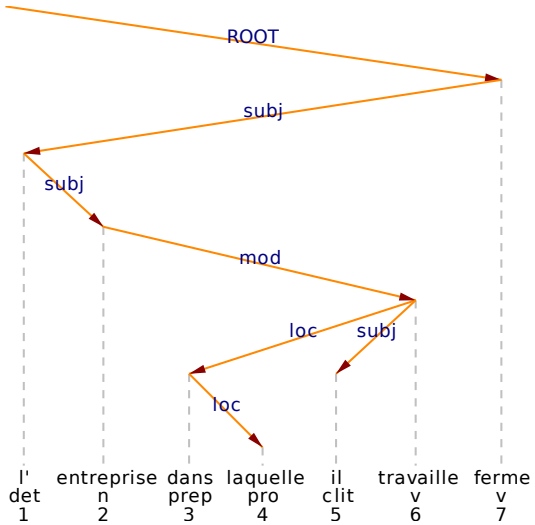
# Les Grammaires d'Interaction (GI)

- Les Grammaires d'Interaction (GI) [GP09] sont un formalisme grammatical qui s'appuie sur les trois notions suivantes :
  - description d'arbre
  - polarités (Grammaires Catégorielles)
  - superposition de structures (HPSG)
- Les GI sont lexicalisées

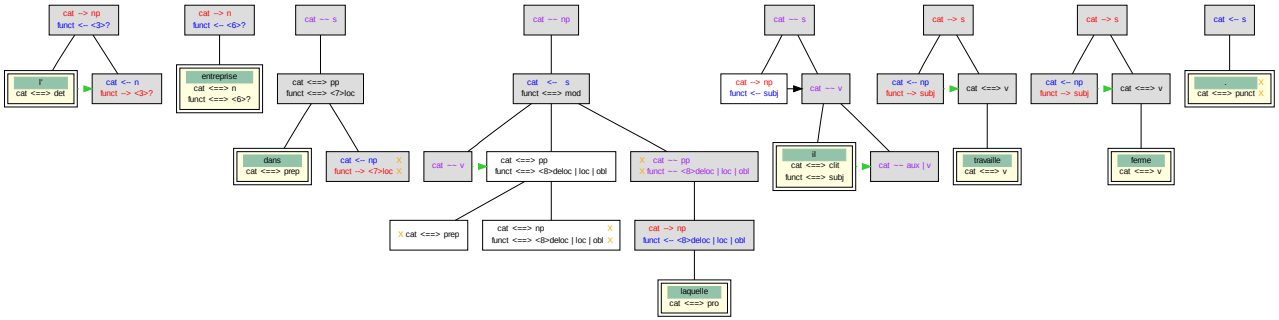
# Analyse dans les GI

- L'analyse dans les GI produit des solutions sous forme d'arbres syntagmatiques et de structures de dépendances
- Exemple pour la phrase *"L'entreprise dans laquelle il travaille ferme."* :

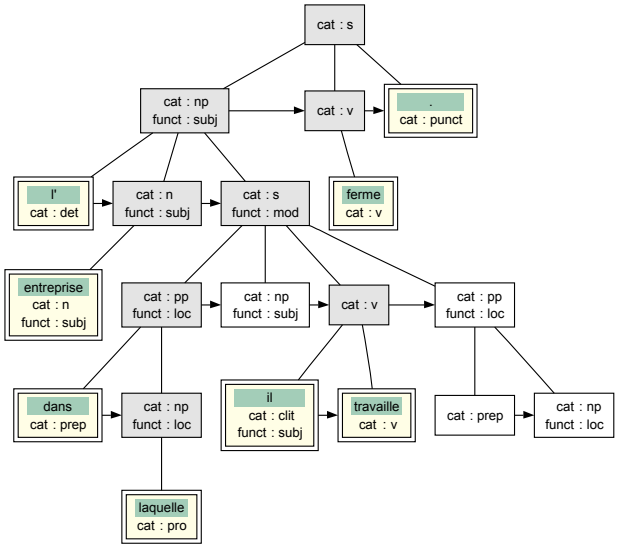


# Analyse dans les GI



Descriptions d'arbre produites par XMG

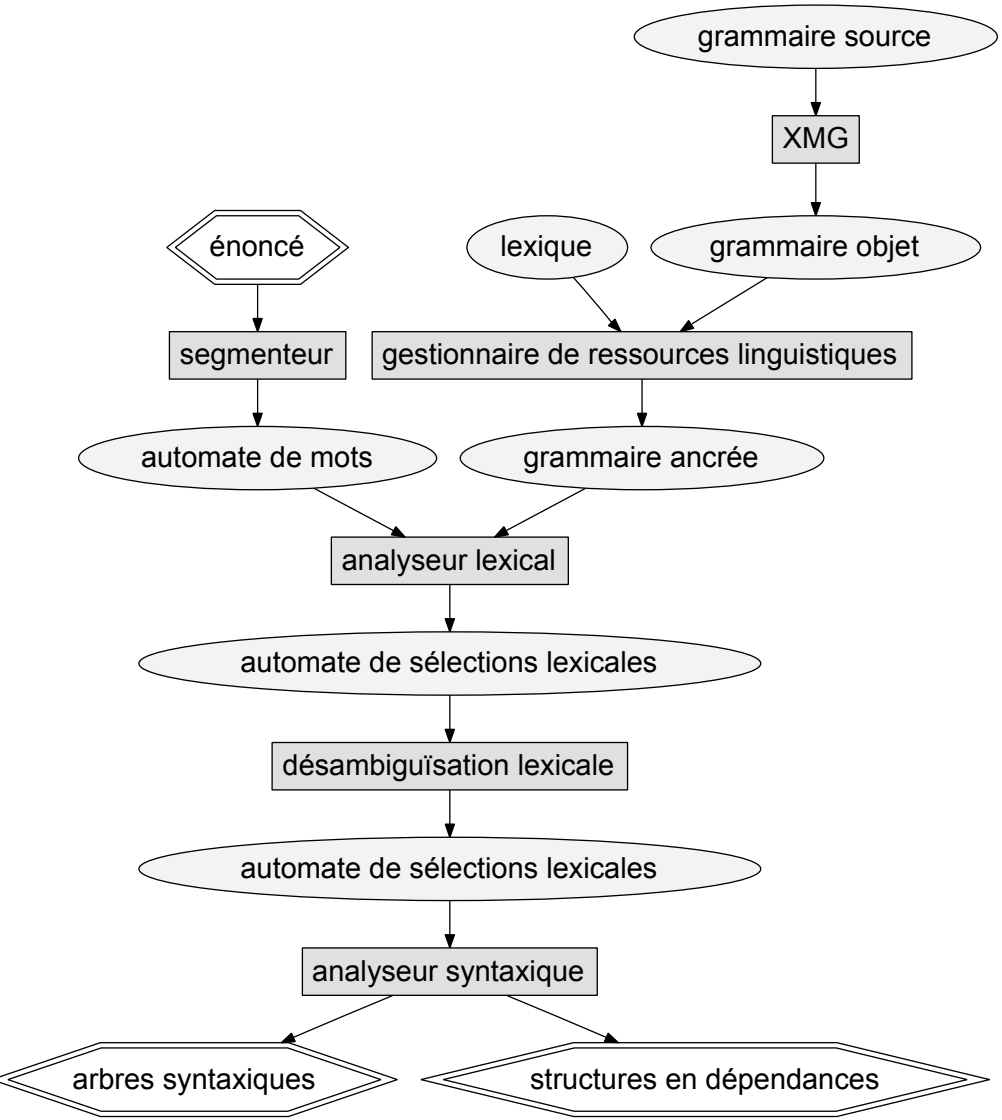



Analyse dans LEOPAR

# Plan



# Architecture

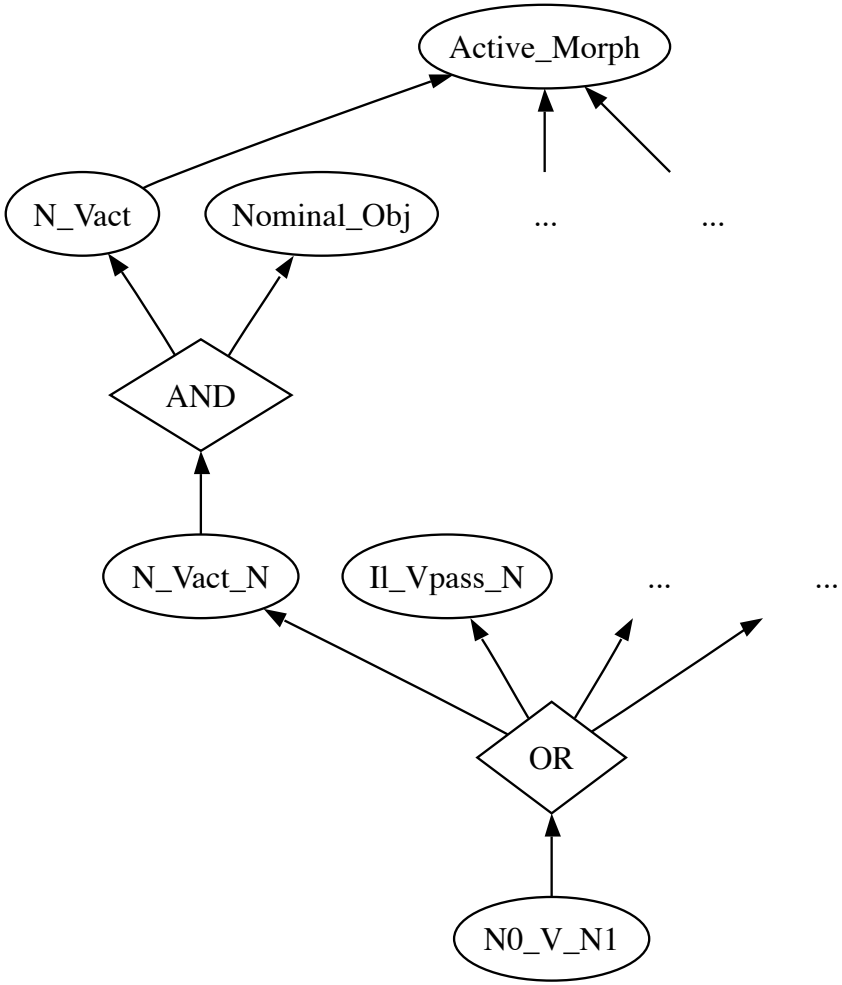



- XMG
  `http://sourcesup.cru.fr/xmg`
- LEOPAR
  `http://leopar.loria.fr`

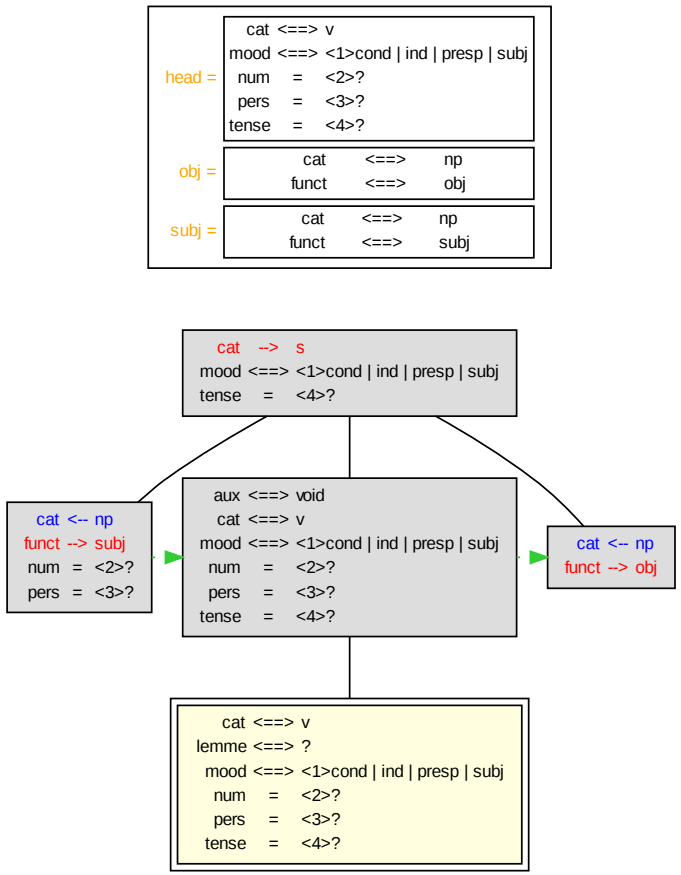
# Plan



## Diagramme de classe partiel définissant la classe `N0_V_N1`




## Exemple de description d'arbre appartenant à la classe `N0_V_N1`



# XMG

- XMG [DLRP04] compile une *grammaire source* (écrite par un humain) vers une *grammaire objet* (utilisée par un système de TAL)
- Propriétés de XMG :
  - Héritage multiple (conjonction ou disjonction de classes).
  - Gestion du nommage des variables
  - Multi-dimensions :
    - Syntaxe
    - Interface avec le lexique
    - Sémantique
    - ...
- XMG est utilisé pour le développement de grammaires TAG et IG, et est conçu pour s'étendre à d'autres formalismes basés sur les arbres

# Plan



# Motivations

- Un mot $\simeq$ 7 entrées lexicales
- Pour une phrase de $n$ mots, il y a $7^n$ étiquetages
- On cherche des méthodes globales et exactes pour éliminer un maximum d'étiquetages
- Supertagging :
  - Abstraire le formalisme dans un formalisme plus simple
  - Analyser les étiquetages dans le formalisme simplifié
  - Garder seulement les étiquetages reconnues

# Désambiguïsation lexicale avec les polarités

- On oublie les informations de structure des objets lexicaux pour ne garder que les polarités
- Une sélection lexicale ne peut être analysé si ses polarités ne peuvent être toutes saturées
- Idée : Vérifier dans chaque sélection lexicale si le bilan des polarités est équilibrée
- Cette vérification peut se faire sous la forme compact d'un automate à états finis

# Évaluation

- Exemples de résultats :

| Énoncé | longueur | étiquetages | étiquetages gardés | ratio |
|---|---|---|---|---|
| Il raté la réunion. | 5 | 74 920 | 42 | $1743^{-1}$ |
| L'entreprise dans laquelle il travaille ferme. | 7 | 955 417 680 | 16128 | $58910^{-1}$ |
| On lui a fait comprendre qu'il aurait dû partir. | 9 | $\simeq$ 170 000 000 000 000 | 84 517 110 | $2011427^{-1}$ |

- Plus la phrase est longue, plus la proportion d'étiquetages gardés est petite
- Mais le nombre d'étiquetages est toujours exponentiel par rapport à la longueur de la phrase

# Désambiguïsation lexicale avec les dépendances

- On considère pour chaque polarités d'un objet lexical les objets lexicaux qui peuvent la saturer par la gauche et ceux qui peuvent le saturer par la droite
- On appelle ces objets lexicaux *les compagnons* de la polarité considérée
- On garde tous les étiquetages où chaque polarité trouve au moins un compagnon

# Évaluation

*"L'entreprise dans laquelle il travaille ferme."*

| **Méthode de désambiguïsation** | **Nombre de sélections lexicales** | **temps** |
|---|---|---|
| sans désambiguïsation | 955 417 680 | 0s |
| polarité | 16 128 | 1,60s |
| dépendances | 744 | 3,48s |
| polarité + dépendances | 101 | 0,73s |

- En combinant les deux stratégies, on passe d'une ambiguïté par mots de 6,31 à 1,41

# Plan



# Analyse syntaxique

- Il existe trois stratégies d'analyse pour les GI :
  - CYK
  - Earley
  - Shift/Reduce
- Pour l'instant seules CYK et Shift/Reduce sont mises en œuvre dans LEOPAR
- Shift/Reduce :
  - Parcours de la sélection lexicale de gauche à droite
  - Shift : Ajouter une description d'arbre dans la pile
  - Reduce : Fusionner deux nœuds des descriptions d'arbre dans la pile pour les saturer
  - Reduce s'applique quand le nombre de polarités positives et négatives dépassent un seuil donné

# Plan



# TSNLP

- La grammaire actuelle du français contient 2380 arbres élémentaires (issus de 373 modules de la *grammaire source*)
- TSNLP du français (Test Suite for Natural Language Processing) :
  - $\simeq$ 1300 phrases grammaticales
  - $\simeq$ 1600 phrases non grammaticales
- Résultats :
  - 88% des phrases grammaticales sont analysées correctement
  - 85% des phrases non grammaticales sont rejetées
- Ces résultats assurent :
  - Large couverture
  - Bonne précision
  - Surgénération limitée

# Passage

- Nécessité d'évaluer sur de grands volumes
- LEOPAR participe à la campagne Passage
- Deux problèmes se posent :
  - Pas de robustesse
  - Ambiguïté lexicale trop grande pour les phrases supérieures à 20 mots
- Cette évaluation est un premier pas pour mesurer les évolutions futures de LEOPAR

# Plan



# Conclusions et perspectives

- Actuellement :
  - Grammaire du français fine et à large couverture
  - Outil complet pour l'analyse profond de phrases inférieures à 20 mots
- À l'avenir :
  - Mise en œuvre de méthodes d'analyse robustes
  - Introduction de statistiques pour diminuer les ambiguïtés
  - Intégration de la sémantique
  - Multilinguisme

# Bibliographie


D. Duchier, J. Le Roux, and Y. Parmentier.
The metagrammar compiler : A NLP Application with a Multi-paradigm Architecture.
In *Second International Mozart/Oz Conference - MOZ 2004, Charleroi, Belgium*, 2004.

B. Guillaume and G. Perrier.
Interaction Grammars.
*Research on Language and Computation*, 2009.
A paraître. Un rapport préliminaire est disponible à l'URL http ://www.loria.fr/ guillaum/publications/RR-6621.pdf.

# Questions ?