

# Using NOOJ's to parse constituents in the french PASSAGE corpus

**Christine Fay-Varnier, Qiuyue Li,  
Azim Roussanaly**  
LORIA, équipe TALARIS  
Nancy, France

# Summary

- Context and project motivation
- Resources
- Method
- Results
- Perspectives

# Context and project motivation

## PASSAGE Evaluation Campaign (Produire des Annotations Syntaxiques à Grande Echelle)

twofold main motivations

- to **improve the accuracy and robustness of existing French parsers** by using them on **large scale corpora** (several millions of words)
- to exploit the resulting syntactic annotations to **create richer and more extensive linguistic resources**

# Evaluation Campaign

- For each sentence in the reference corpus the submitted parsers must provide
  - the set of the major constituents occurring in that sentence
  - and the grammatical relations between these constituents
- Our approach follows a two step procedure:
  - inter words relations are extracted from the derivation trees obtained with a deep syntactic parser based on the LLP2 tree adjoining grammar developed by the LORIA team
  - these relations are projected onto the constituents.

# The resources

- A syntactic annotations reference Guide: PEAS
- LLP2 analyser lexicon
- Corpus in various domains: newspaper, litterature, mail, oral, medical (>40000 sentences)
- The annotated corpus
- A test platform

# The annotations reference guide

- 6 constituent types: NV, PV, GN, GP, GA, GR

- Rules:

« Les **groupes prépositionnels** sont constitués d'une préposition et du GN qu'elle introduit ou d'un déterminant et d'une préposition contractés (*du, aux, des*) avec le GN introduit. On forme aussi un GP en cas de préposition suivie d'un adverbe. On considère également comme des GP les pronoms qui remplacent des GP, comme *dont, où,...* »

- Annotated examples:

« *la porte <GP> de la chambre </GP> fermée <GP> à clef </GP><GP> à l'intérieur </GP> , les volets <GP> de l'unique fenêtre </GP> fermés , eux aussi , <GP> à l'intérieur </GP> , et , <GP> par-dessus les volets </GP> ,... »*

# The Lexicon

- 17 categories

...  
partir,.NCFF  
partira,.CJ  
partirai,.CJ  
partirais,.CJ  
partirait,.CJ  
partirent,.CJ  
partirez,.CJ  
partirons,.CJ  
partis,.CJ  
partis,.PAR  
partis,.SBC  
partisan,.ADJ  
partisan,.SBC  
Partisans,.ADJ  
Partisans,.SBC  
partisans,.ADJ  
partisans,.SBC

- Locutions Dictionary

...  
à outrance,.ADV  
à part,.ADV  
à partir d',.PREP  
à partir de,.PREP  
à partir des,.PREP

# The corpus

- Xml format:

« Cet outil sera réalisé à partir des technologies multilingues développées dans le cadre du projet européen Emir»

```
<E ID="E166">
  <F ID="E166F1">Cet</F>
  <F ID="E166F2">outil</F>
  <F ID="E166F3">sera</F>
  <F ID="E166F4">réalisé</F>
  <F ID="E166F5">à partir des</F>
  <F ID="E166F6">technologies</F>
  <F ID="E166F7">multilingues</F>
  <F ID="E166F8">développées</F>
  ...
</E>
```



# The annotated corpus

```
<E id="E166">
  <Groupe type="GN" id="E166G1">
    <F id="E166F1">Cet</F>
    <F id="E166F2">outil</F>
  </Groupe>
  <Groupe type="NV" id="E166G2">
    <F id="E166F3">sera</F>
  </Groupe>
  <Groupe type="NV" id="E166G3">
    <F id="E166F4">réalisé</F>
  </Groupe>
  <Groupe type="GP" id="E166G4">
    <F id="E166F5">à_partir_des</F>
    <F id="E166F6">technologies</F>
  </Groupe>
  <Groupe type="GA" id="E166G5">
    <F id="E166F7">multilingues</F>
  </Groupe>
  <Groupe type="NV" id="E166G6">
    <F id="E166F8">développées</F>
  </Groupe>
  ...
</E>
```

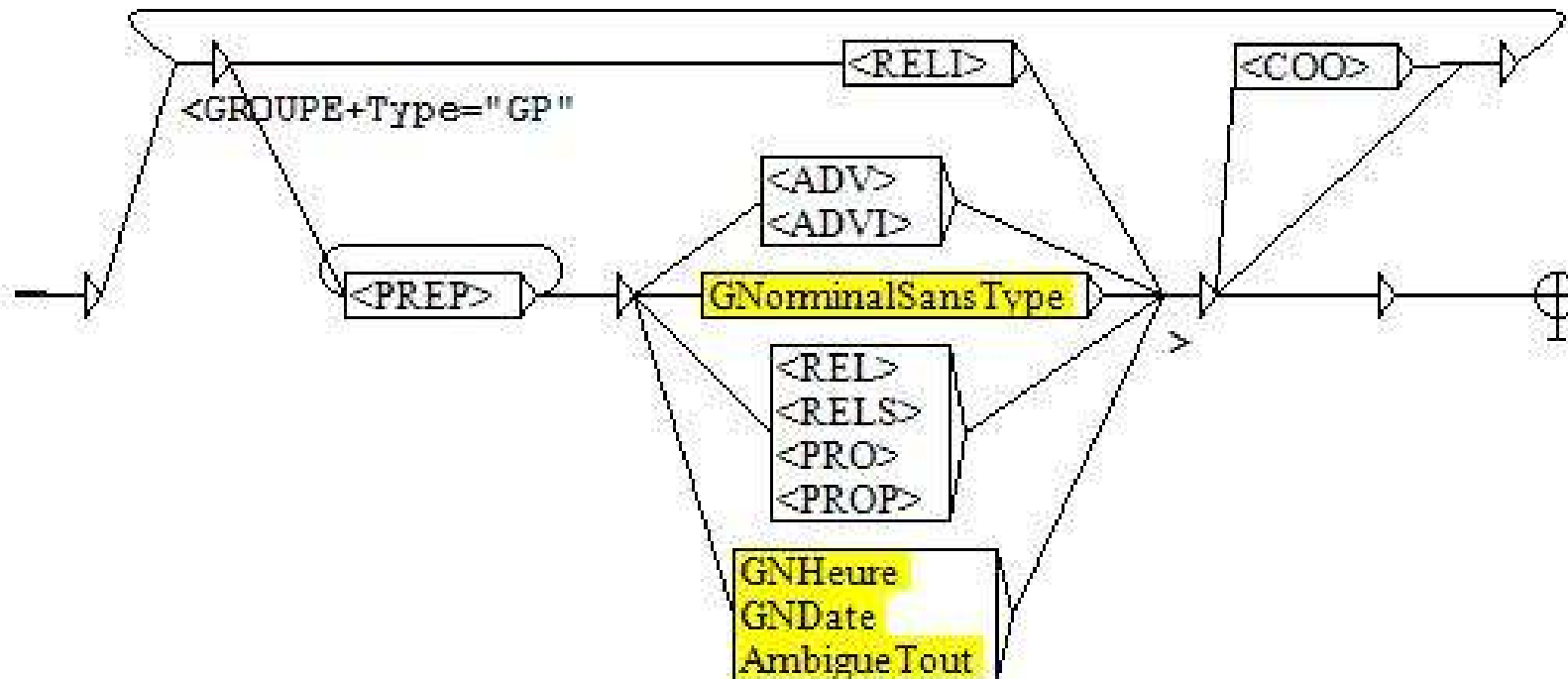
# Pragmatic methodology

- Developing an initial set of FST from PEAS/PASSAGE reference annotation
- Testing the rules set on all available annotated corpora
- Correctly according the rules to the results obtained
- Trying to resolve some ambiguities
- Itering the process

# GP FST

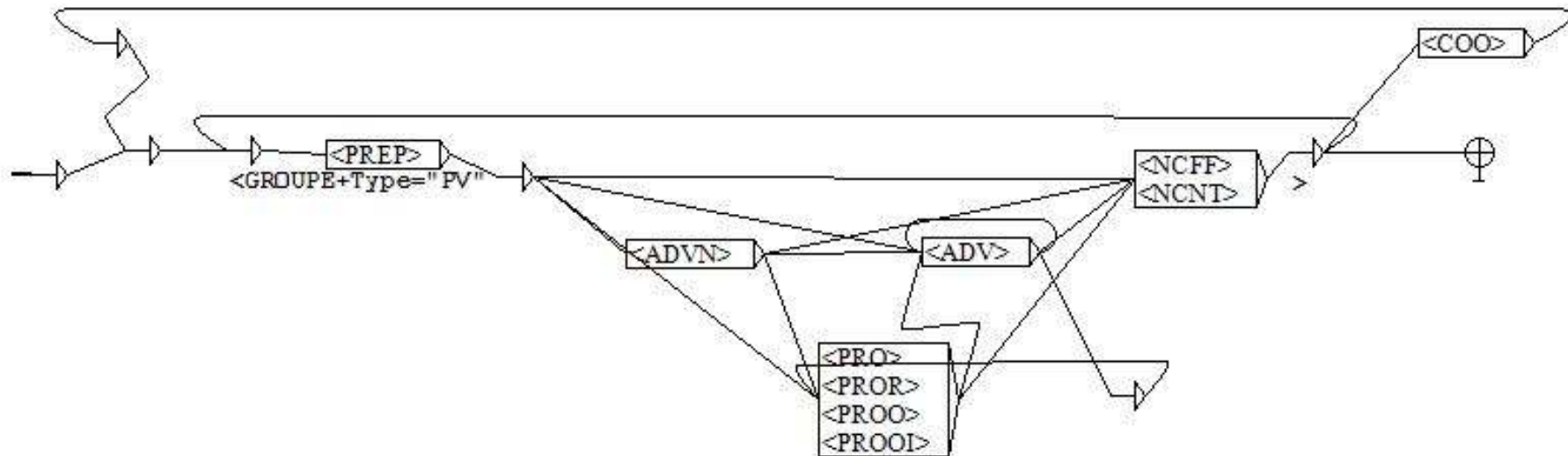
« **<GP>A quoi</GP> servent les ressources linguistiques** »

« **...des lexiques spécialisés <GP>à partir d'un ensemble</GP> de textes techniques** »



# PV FST

« Leur qualité repose sur les données linguistiques utilisées <GP>pour entraîner</GP> le système. »

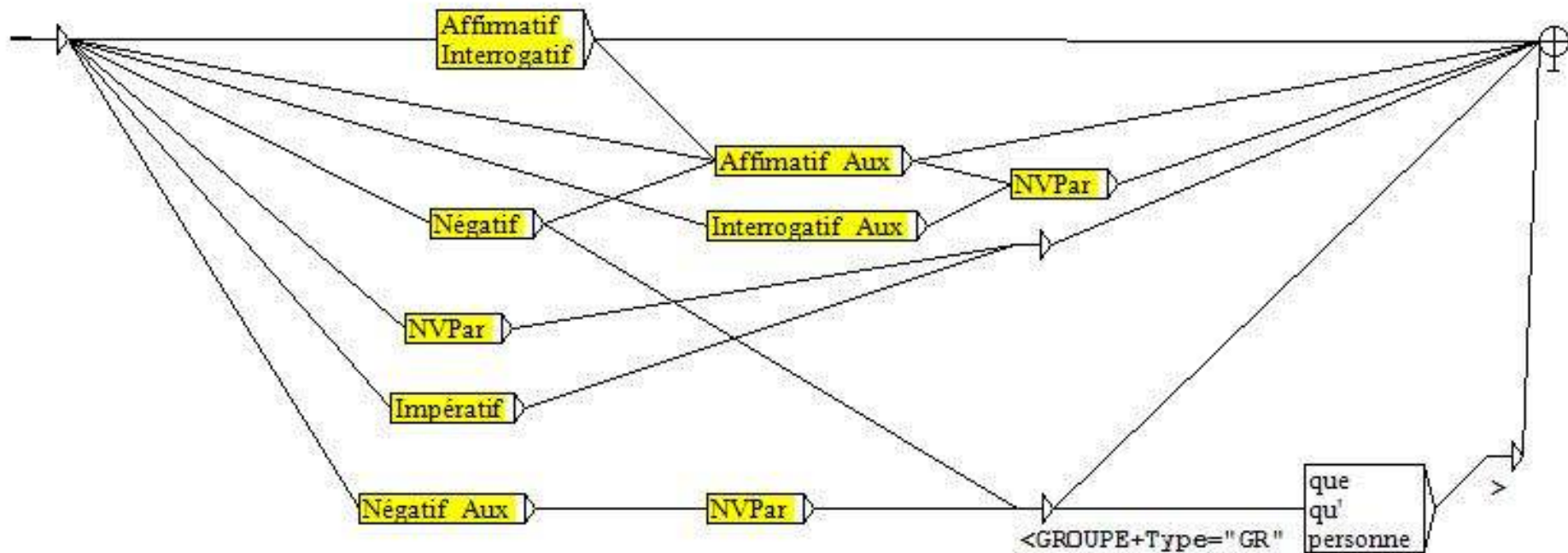


# Accord the rules

- The errors:
  - mainly caused by system's bad choice of ambiguity rules
  - can be reduced by:
    1. Sophisticating categories and in particular specializing some of them  
« *the PRO category can be divided into sub-categories : PROS, PROO, PROP, PROR, etc.* »
    2. Using the context to disambiguate.

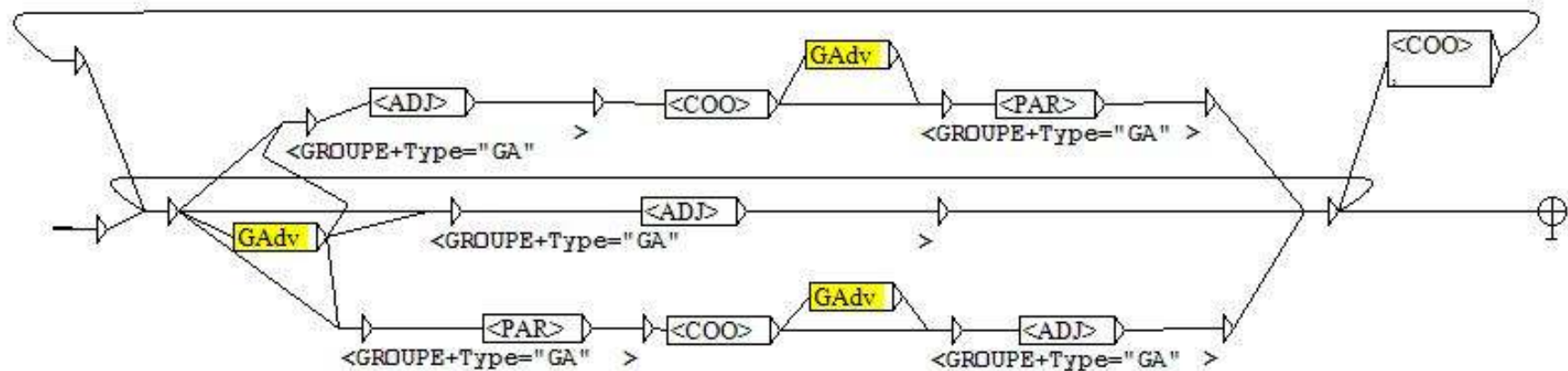
# Local rules of disambiguation

« il ne s'agissait <GR>que</GR> d'une visite de travail »



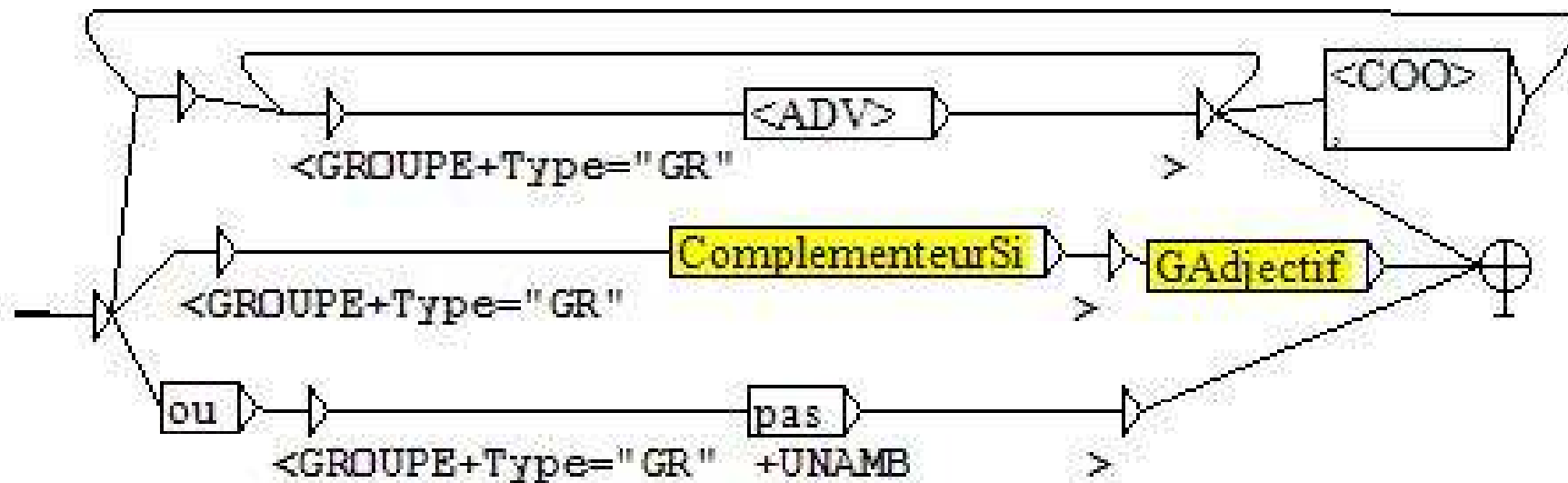
# Local rules of disambiguation

« "les systèmes de traitement du langage **<GA> oral </GA>** ou **<GA> écrits </GA>** »"



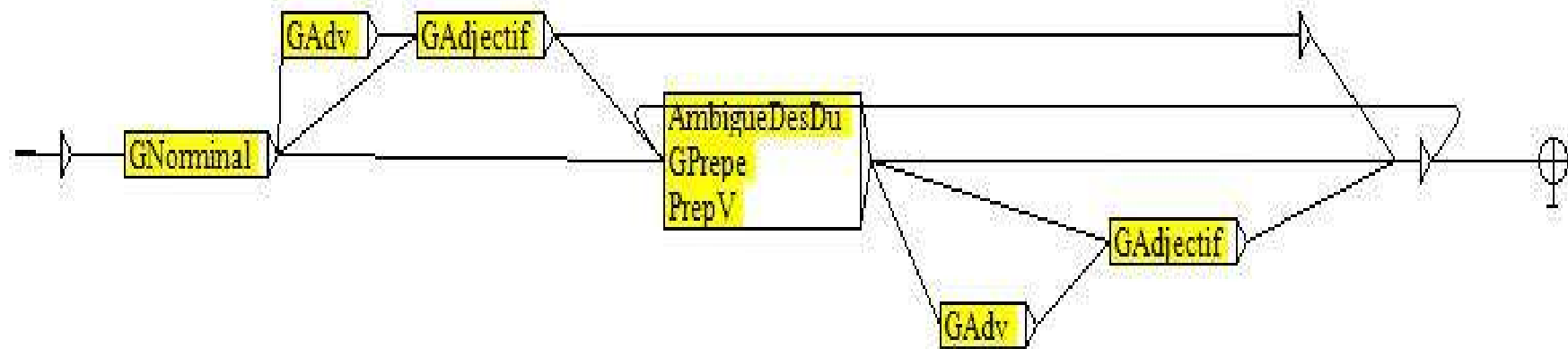
# Local rules of disambiguation

«...l'exploitation des ressources linguistiques intégrées ou <GR> pas</GR> dans des produits... »



# Larger context rules of disambiguation

«... un embargo total sur les marchandises à destination des zones sous contrôle...»



# Results

- Evaluation corpus : ~5000 sentences
- Results of december with INTEX

GN			NV			GA			GR			GP			PV			All		
P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
87.5	90.7	89.0	92.0	97.1	94.4	87.4	83.2	85.3	88.3	88.3	85.7	96.3	95.1	95.7	97.7	91.5	94.5	91.5	92.2	91.9



# Some difficulties

- Language ambiguity
  - Constituents beginning with « du » or « des »
  - Past participles
    - « *Des lexiques <GA>spécialisés</GA> sont utilisés... »*
    - « *Des centres agréés(NV) trop peu nombreux, ... »*
  - Ambiguïté du « que »: CONJ or REL
    - « *que <NV>je place</NV> »*
  - Etc.
- Miss expertise on NOOJ ...
  - Traitment of xml
  - Gestion of priority
  - Etc.

# Perpective

- Project to use syntactic and/or semantic informations to improve the accuracy of a speech recognizer
- No reasonable to do deep analysis →  
Chunker