Probabiliser les Grammaires de Propriétés

Philippe Blache & Stéphane Rauzy

Laboratoire Parole et Langage
CNRS & Aix-Marseille Université

Introduction

- L'analyse en Grammaire de Propriétés
- Facteurs de complexité
- Quelles informations pour guider l'analyse ?
- Comment intégrer les probabilités ?

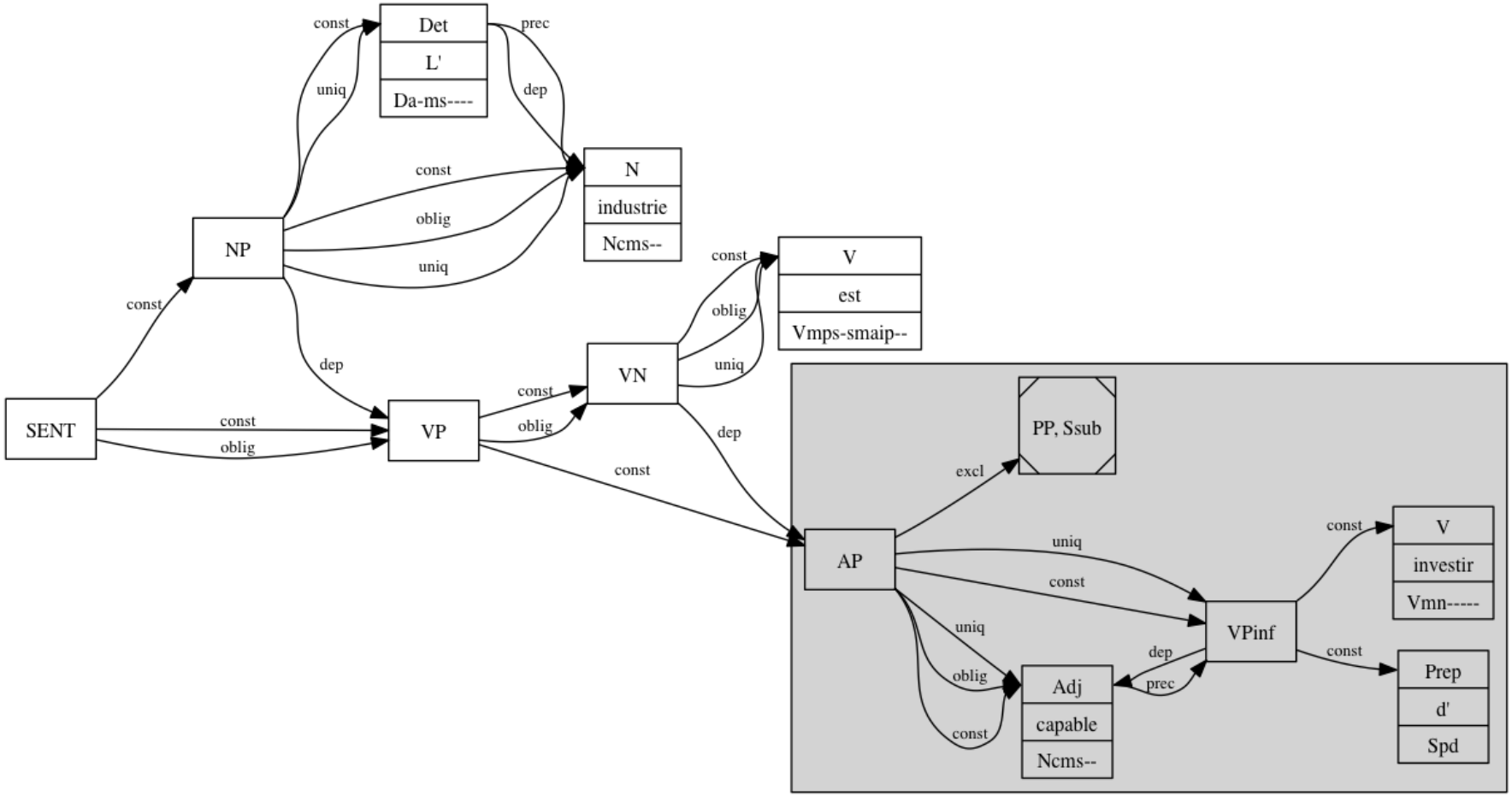
Les contraintes en GP

Obligation	$A : \Delta B$	au moins un B
Unicité	$A : B!$	au plus un B
Linéarité	$A : B \prec C$	B précède C
Implication	$A : B \Rightarrow C$	si $\exists B$, alors $\exists C$
Exclusion	$A : B \not\Rightarrow C$	pas de B et C simultanément
Constituance	$A : S?$	les descendants $\in S$
Adjacence	$A : B \oplus C$	A est adjacent à C
Dépendance	$A : B \rightsquigarrow C$	B dépend de C

Exemple : le SA

Constituance	AP: $\{$AdP, A, VPinf, PP, Ssub, AP, NP$\}$?
Linéarité	AP: A $\prec$ $\{$VPinf, Ssub, PP, NP, AP$\}$ AP: AdP $\prec$ $\{$A, Ssub, PP$\}$ AP: AP $\prec$ $\{$A, AdP$\}$ AP: PP $\prec$ $\{$Ssub$\}$
Dépendance	AP: $\{$AdP, VPinf, PP, Ssub, NP$\}$ $\rightsquigarrow$ A
Unicité	AP: $\{$A, VPinf, Ssub$\}$!
Obligation	AP: Δ A
Exclusion	AP: VPinf $\not\Leftrightarrow$ $\{$PP, Ssub$\}$

Exemple de graphe de contraintes



Graphe des propriétés satisfaites pour "*L'industrie est capable d'investir.*"

Des contraintes multidimensionnelles

- Relation contour/groupe prosodique:

$$\texttt{Min\_Rising}_{[proso]} \ : \ \{\texttt{AP}_{[proso]}\}$$

Les syntagmes accentuels se terminent par des contours à montée mineure

- Relation groupe syntaxique/contour :

$$\texttt{NV}_{[synt]} : \{\texttt{V}_{[synt]}, \texttt{Min\_Rising}_{[proso]}\}$$

Les noyaux verbaux sont réalisés par des contours à montée mineure

- Backchannels :

$$\begin{aligned} \texttt{IP}_{[proso]} &\prec \texttt{VocBC}_{[disc]} \\ \texttt{VocBC}_{[disc]} &\prec \{\texttt{RF}, \texttt{RT}\}_{[proso]} \\ \texttt{VocBC}_{[disc]} &\prec \texttt{TCU\_f}_{[disc]} \end{aligned}$$

Les backchannels apparaissent après des IP, réalisés par des contours montant-descendants ou montées terminales, et après des tours conversationnels complets

Des contraintes multidimensionnelles

- Relation contour/groupe prosodique:

$$\texttt{Min\_Rising}_{[proso]} \ : \ \{\texttt{AP}_{[proso]}\}$$

Les syntagmes accentuels se terminent par des contours à montée mineure

- Relation groupe syntaxique/contour :

$$\text{NV}_{[synt]} : \{\text{V}_{[synt]}, \text{Min\_Rising}_{[proso]}\}$$

Les noyaux verbaux sont réalisés par des contours à montée mineure

- Backchannels :

$$\begin{aligned}
&\texttt{IP}_{[proso]} \prec \texttt{VocBC}_{[disc]} \\
&\texttt{VocBC}_{[disc]} \prec \{\texttt{RF}, \ \texttt{RT}\}_{[proso]} \\
&\texttt{VocBC}_{[disc]} \prec \texttt{TCU_f}_{[disc]}
\end{aligned}$$

Les backchannels apparaissent après des IP, réalisés par des contours montant-descendants ou montées terminales, et après des tours conversationnels complets

Des contraintes multidimensionnelles

- Relation contour/groupe prosodique:

$$\texttt{Min\_Rising}_{[proso]} \ : \ \{\texttt{AP}_{[proso]}\}$$

Les syntagmes accentuels se terminent par des contours à montée mineure

- Relation groupe syntaxique/contour :

$$\text{NV}_{[synt]} : \{\text{V}_{[synt]}, \text{Min\_Rising}_{[proso]}\}$$

Les noyaux verbaux sont réalisés par des contours à montée mineure

- Backchannels :

$$\begin{aligned} \texttt{IP}_{[proso]} &\prec \texttt{VocBC}_{[disc]} \\ \texttt{VocBC}_{[disc]} &\prec \{\texttt{RF}, \ \texttt{RT}\}_{[proso]} \\ \texttt{VocBC}_{[disc]} &\prec \texttt{TCU\_f}_{[disc]} \end{aligned}$$

Les backchannels apparaissent après des IP, réalisés par des contours montant-descendants ou montées terminales, et après des tours conversationnels complets

L'analyse en GP

1. Input : ensemble $\mathcal{C}$ des catégorisations possibles
2. Génération de l'ensemble $\mathcal{A}$ des affectations (ensemble des sous-ensembles de $\mathcal{C}$)
3. Construction des caractérisations de $\mathcal{A}$: évaluation des propriétés décrivant les relations entre catégories pour chaque affectation
4. Inférence de nouvelles catégories : si une caractérisation correspond à une catégorie, on l'ajoute à $\mathcal{C}$
5. Recherche d'une solution : partition de $\mathcal{A}$
   - qui couvre la totalité de l'input
   - dans lequel chaque mot correspond à une étiquette unique
   - qui satisfait un nombre suffisant de propriétés

L'analyse en GP

1. Input : ensemble $\mathcal{C}$ des catégorisations possibles
2. Génération de l'ensemble $\mathcal{A}$ des affectations (ensemble des sous-ensembles de $\mathcal{C}$)
3. Construction des caractérisations de $\mathcal{A}$: évaluation des propriétés décrivant les relations entre catégories pour chaque affectation
4. Inférence de nouvelles catégories : si une caractérisation correspond à une catégorie, on l'ajoute à $\mathcal{C}$
5. Recherche d'une solution : partition de $\mathcal{A}$
   - qui couvre la totalité de l'input
   - dans lequel chaque mot correspond à une étiquette unique
   - qui satisfait un nombre suffisant de propriétés

L'analyse en GP

1. Input : ensemble $\mathcal{C}$ des catégorisations possibles
2. Génération de l'ensemble $\mathcal{A}$ des affectations (ensemble des sous-ensembles de $\mathcal{C}$)
3. Construction des caractérisations de $\mathcal{A}$: évaluation des propriétés décrivant les relations entre catégories pour chaque affectation
4. Inférence de nouvelles catégories : si une caractérisation correspond à une catégorie, on l'ajoute à $\mathcal{C}$
5. Recherche d'une solution : partition de $\mathcal{A}$
   - qui couvre la totalité de l'input
   - dans lequel chaque mot correspond à une étiquette unique
   - qui satisfait un nombre suffisant de propriétés

L'analyse en GP

1. Input : ensemble $\mathcal{C}$ des catégorisations possibles
2. Génération de l'ensemble $\mathcal{A}$ des affectations (ensemble des sous-ensembles de $\mathcal{C}$)
3. Construction des caractérisations de $\mathcal{A}$: évaluation des propriétés décrivant les relations entre catégories pour chaque affectation
4. Inférence de nouvelles catégories : si une caractérisation correspond à une catégorie, on l'ajoute à $\mathcal{C}$
5. Recherche d'une solution : partition de $\mathcal{A}$
   - qui couvre la totalité de l'input
   - dans lequel chaque mot correspond à une étiquette unique
   - qui satisfait un nombre suffisant de propriétés

L'analyse en GP

1. Input : ensemble $\mathcal{C}$ des catégorisations possibles
2. Génération de l'ensemble $\mathcal{A}$ des affectations (ensemble des sous-ensembles de $\mathcal{C}$)
3. Construction des caractérisations de $\mathcal{A}$: évaluation des propriétés décrivant les relations entre catégories pour chaque affectation
4. Inférence de nouvelles catégories : si une caractérisation correspond à une catégorie, on l'ajoute à $\mathcal{C}$
5. Recherche d'une solution : partition de $\mathcal{A}$
   - qui couvre la totalité de l'input
   - dans lequel chaque mot correspond à une étiquette unique
   - qui satisfait un nombre suffisant de propriétés

Exemple

La	*principale*	*activité*	*du*	*domaine*	...
Det	Adj	N	Prep	N	
Pro	N				
N					

- Affectations $\mathcal{A} = \{\texttt{Det}\}; \{\texttt{Adj}\}; \{\texttt{Det, Adj}\}; \{\texttt{Det, Adj, N}\}; \{\texttt{Pro, Adj}\}, \{\texttt{N, Prep, N}\}$, etc.
- Caractérisation de $\{\text{Det}\}$: $\mathcal{C}^+ = \{NP : Det!\}$ $\mathcal{C}^- = \{NP : Det \Rightarrow N\}$
  Caractérisation de $\{\text{Adj}\}$: $\mathcal{C}^+ = \{AP : Adj!; AP : \Delta Adj\}$
- Inférence de AP

Facteurs de complexité

- Complexité classique des formalismes DI/PL : dépend du nombre de contraintes linéaires
- Taille de l'ensemble des catégories :
  - Croissant au fur et à mesure de l'analyse
  - Ensemble potentiellement grand : prise en compte de toutes les dimensions
- Taille de l'espace de recherche : ensemble des affectations générées
- Recherche d'une partition couvrant l'input
- Evaluation des propriétés : fonction de la taille de la grammaire

Données initiales : treebank enrichi

```
<category label="NP" features="NP:SUJ" node_index="0:16:1">
   <category label="Det" features="Da-ms----" node_index="0:16:2" form="Le
   <category label="Noun" features="Ncms--" node_index="0:16:3" form="text
   <category label="AP" features="AP" node_index="0:16:4">
      <category label="Adj" features="Af-ms-" node_index="0:16:5" form="mê
   </category>
   <characterization>
      <property type="lin" source="0:16:2" target="0:16:3" sat="p"/>
      <property type="lin" source="0:16:2" target="0:16:4" sat="p"/>
      <property type="req" source="0:16:3" target="0:16:2" sat="p"/>
      <property type="dep" source="0:16:2" target="0:16:3" sat="p"/>
      <property type="dep" source="0:16:4" target="0:16:3" sat="p"/>
      <property type="oblig" source="0:16:1" target="0:16:3" sat="p"/>
      <property type="uniq" source="0:16:1" target="0:16:2" sat="p"/>
   </characterization>
</category>
```

Affectations caractérisées

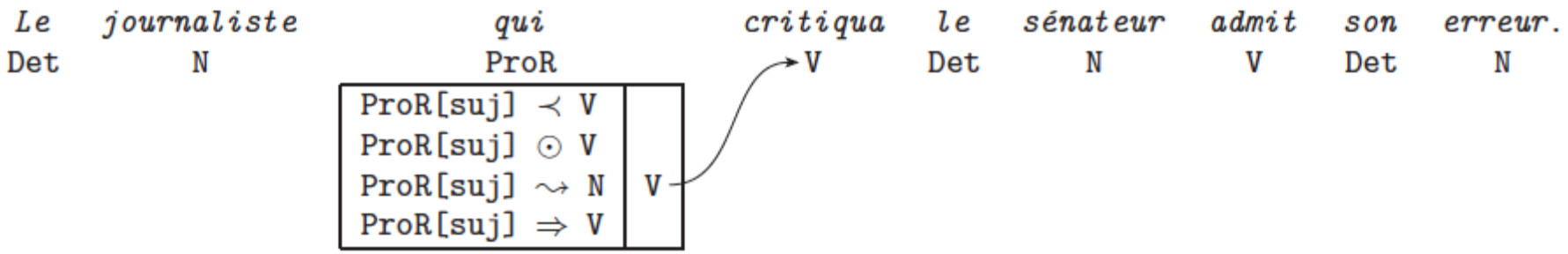
Propriétés	Réalisations				
	{Det,N}	{Det,N,AP}	{Det,N,PP}	{Det,N,AP,PP}	{N,NP}
[1] Det $\prec$ N	x	x	x	x	
[2] Det $\prec$ AP		x		x	
[3] Det $\prec$ PP			x	x	
[4] Det $\prec$ NP					x
[5] N $\prec$ PP			x	x	
[6] Oblig(N)	x	x	x	x	x
[7] Det $\rightsquigarrow$ N	x	x	x	x	
[8] PP $\rightsquigarrow$ N			x	x	
[9] AP $\rightsquigarrow$ N		x		x	
[10] Det $\Rightarrow$ N	x	x	x	x	x
[11] PP $\Rightarrow$ N			x	x	
[12] AP $\Rightarrow$ N		x		x	
[13] Unic(Det)	x	x	x	x	
[14] Unic(N)	x	x	x	x	
[15] N $\not\Leftrightarrow$ Pro	x	x	x	x	x
[16] N $\not\Leftrightarrow$ Npro	x	x	x	x	x
[17] AP $\not\Leftrightarrow$ Pro		x		x	
...					

Prédiction

- En CFG : étant donné un ensemble de catégories (contexte), on prédit une transition (nouvelle catégorie)
- En GP : étant donné un ensemble de contraintes, on peut :
  - prédire une transition
  - prédire une caractérisation
  - prédire une projection (inférer une catégorie supérieure)

Prédiction de transition : l'activation

Un ensemble de contraintes active un certain nombre de catégories :



Calcul de *l'Activation*

$$A(c_j) = \frac{n}{N} * \sum W_{c_i}^{c_j} \quad (1)$$

Contrôle dans le choix de l'affectation

Affectation	$\{\texttt{Det},\texttt{N},\texttt{AP}\}$	$\{\texttt{Det},\texttt{N},\texttt{PP}\}$
Distribution	0.43	0.2
Caractérisation	$\texttt{Det} \prec \texttt{N}$, $\texttt{Det} \prec \texttt{AP}$, $\Delta\texttt{N}$, $\texttt{Det} \rightsquigarrow \texttt{N}$, $\texttt{AP} \rightsquigarrow \texttt{N}$, $\texttt{Det} \Rightarrow \texttt{N}$, $\texttt{AP} \Rightarrow \texttt{N}$, Det!, N!, $\texttt{N} \not\Leftrightarrow \texttt{Pro}$, $\texttt{N} \not\Leftrightarrow \texttt{Npro}$, $\texttt{AP} \not\Leftrightarrow \texttt{Pro}$	$\texttt{Det} \prec \texttt{N}$, $\texttt{Det} \prec \texttt{PP}$, $\texttt{N} \prec \texttt{PP}$, $\Delta\texttt{N}$, $\texttt{Det} \rightsquigarrow \texttt{N}$, $\texttt{PP} \rightsquigarrow \texttt{N}$, $\texttt{Det} \Rightarrow \texttt{N}$, $\texttt{PP} \Rightarrow \texttt{N}$, Det!, N!, $\texttt{N} \not\Leftrightarrow \texttt{Pro}$, $\texttt{N} \not\Leftrightarrow \texttt{Npro}$, $\texttt{PP} \not\Leftrightarrow \texttt{Ssub}$

Caractérisations multidimensionnelles

Exemple :

Marie	je	la	supporte	pas
N	Clit	Pro	V	Adv
NP	S			

Dislocation	
Disl:	NP?, S?, L*H?
Disl:	NP $\prec$ S
Disl:	NP $\oplus$ S
Disl:	NP $\rightsquigarrow$ S//Clit
Disl:	NP $\approx$ L*H

Vocatif	
Voc:	NP?, S?, H*L?
Voc:	NP $\prec$ S
Voc:	NP $\oplus$ S
Voc:	NP $\approx$ H*L

Prédiction de projection

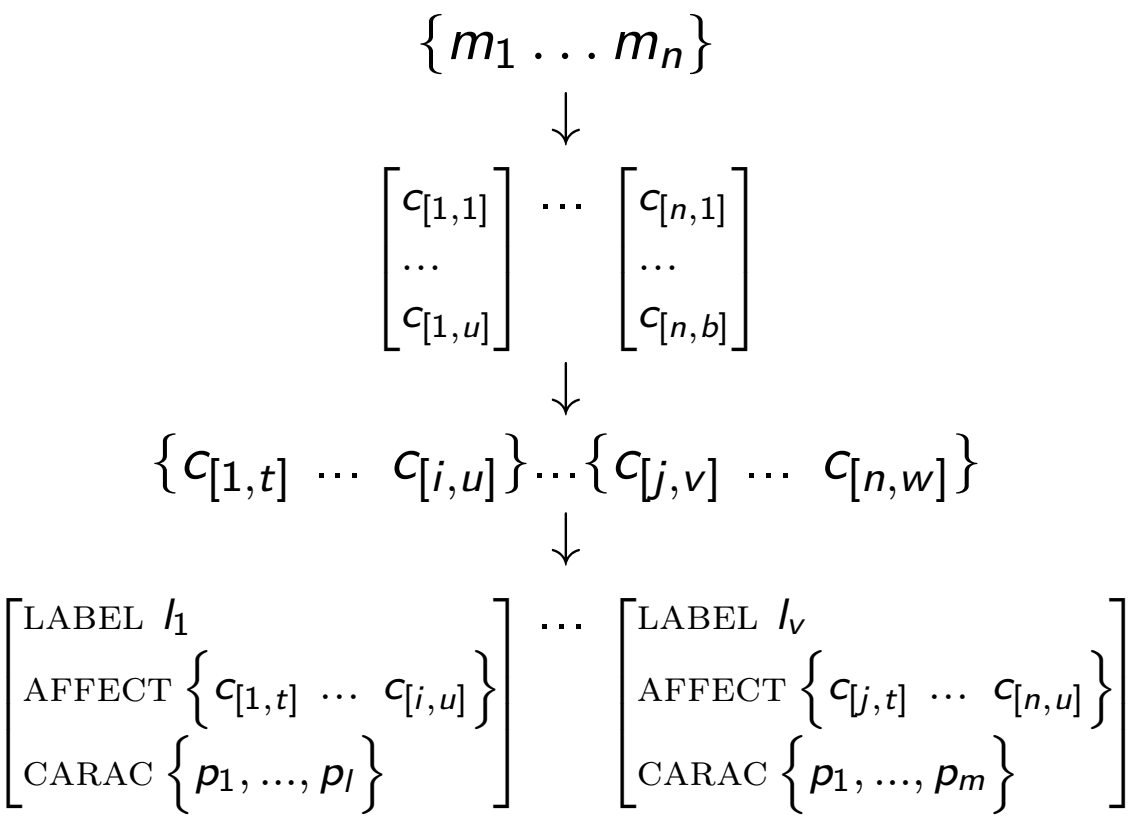
Distribution :

- $\{\text{NP?, S?, NP} \prec \text{S, NP} \approx \text{L*H}\} = 0.7$
- $\{\text{NP?, S?, NP} \prec \text{S, NP} \approx \text{H*L}\} = 0.3$

En rencontrant $\{\text{NP?, S?, NP} \prec \text{S}\}$, on prédit comme structure la plus probable la projection de la `dislocation` en même temps que sa caractérisation complète:

Disl: $\{\text{NP?, S?, L*H?, NP} \oplus \text{S, NP} \prec \text{S, NP} \rightsquigarrow \text{S//Clit, NP} \approx \text{L*H, ...}\}$

Processus



Catégories probabilisées

Affectations probabilisées

Caractérisations probabilisée

Conclusion

- Données observables : treebanks hybrides
- Contrôle
  - Choix des affectations
  - Calcul des caractérisations
- Intérêts
  - Informations partielles
  - Informations sous-spécifiées
  - Données non canoniques