

# Des arbres de dérivation aux forêts de dépendance : un chemin via les forêts partagées

Djamé Seddah    Bertrand Gaiffe

LORIA, Vandœuvre-lès-Nancy, France



TALN 2005, Dourdan, France, 06-10 Juin

## Objectif: construire un graphe de dépendance à partir d'une analyse TAG

A partir d'une analyse TAG, nous voulons :

- ▶ Construire une représentation de la structure syntaxique profonde : avec toutes les relations argumentales présentes (un graphe de dépendance)
- ▶ Avoir toutes les analyses représentées dans une même structure

Possible si nous travaillons directement sur une structure combinant **Arbre dérivé** et **Arbre de dérivation** : **Une forêt partagée**

# Plan

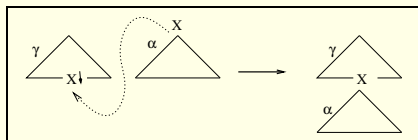
1. On introduit les TAG et le problème de l'arbre de dérivation à travers les phénomènes des verbes à contrôle
2. Puis nous montrons que les informations utiles sont dans les deux structures de représentations qu'offrent les TAG et donc dans une Forêt partagée
3. Et nous présentons finalement notre façon d'utiliser la forêt partagée à travers l'extraction des dérivations TAG

## Un formalisme lexicalisé

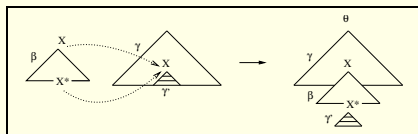
- ▶ Une grammaire constituée d'arbres élémentaires (initiaux et auxiliaires)
- ▶ Chaque arbre est ancré par une unité lexicale
- ▶ Deux opérations : substitution et adjonction
- ▶ Contrairement aux CFG, arbre de dérivation et arbre dérivé ne sont plus isomorphes

## Operations sur les arbres

- **La substitution** est une opération CF de dérivation d'un arbre initial sur le nœud feuille d'un arbre élémentaire



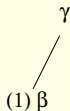
- **L'adjunction** est une opération contextuelle d'insertion d'un arbre auxiliaire au sein d'un arbre élémentaire



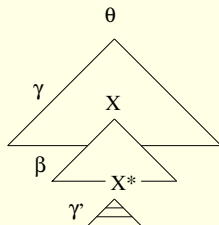
## Arbres de dérivation et arbres dérivés

- ▶ Arbre de dérivation : décrit la construction de l'arbre dérivé (*i.e l'historique des opérations conduisant à la reconnaissance d'un énoncé*)
- ▶ Arbre dérivé : La structure syntaxique d'un énoncé
- ▶ Ex: Soit les arbres  $\gamma$  et  $\beta$ , avec  $\beta$  adjoint sur le nœud 1 de  $\gamma$

Derivation tree

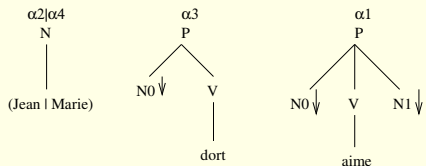


Derived Tree

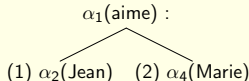


# Illustration sur une phrase simple

Jean aime Marie



Arbre de dérivation:



⇒ L'arbre de dérivation semble être une structure prédicat-argument correcte et peut être vu comme un graphe de dépendances syntaxiques profondes ([Joshi & Rambow, 92])

Structure Prédicat-argument

⇒

$\text{aime-}\alpha_1(\text{Jean-}\alpha_2, \text{Marie-}\alpha_4)$

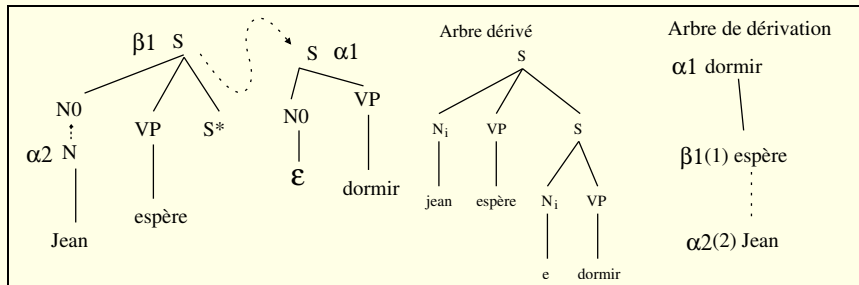
## Est ce que l'arbre de dérivation (DT) est une bonne structure ?

- ▶ Fait : le DT ne tient compte que des dérivations (*le passage d'un arbre vers un autre*)
  - *Conséquence : chaque dérivation doit être un argument sémantique pour que le DT soit une structure sémantique correcte*
- ▶ Fait : L'arbre de dérivation est un arbre et non un graphe
  - *Conséquence : l'arbre de dérivation ne permet pas de rendre compte des co-indications*
- ▶ Fait : Les nœuds de l'arbre de dérivation sont des projections d'arbre
  - *Conséquence : impossibilité d'établir des co-indications entre des nœuds d'arbres autrement que par l'emploi de structure de traits*



## Illustration sur les arbres à contrôle

Problème bien connu: une réalisation syntaxique d'un argument n'est pas présente dans le DT

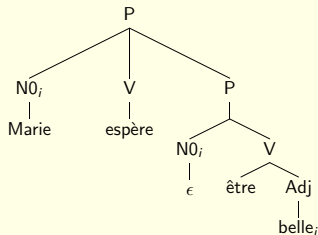


Bien que l'arbre dérivé témoigne d'une coïncidence (ie sous catégorisation ou accord entre les nœuds  $N_i$ , ça n'apparaît pas dans le DT)

Question : Comment créer un lien entre "Jean" et "Dormir" ?

## L'information manquante est dans l'arbre dérivé

Arbre dérivé pour "Marie espère être belle"

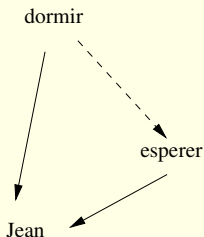


L'accord entre "Marie" et le sujet non réalisé d' "être belle" prouve l'existence d'un lien argumental

On a donc une relation syntaxique visible via l'accord non représentée dans l'arbre de dérivation

## La structure souhaitée : un graphe de dépendance ?

Le lien manquant est visible dans l'arbre dérivé (accord par percolation de traits témoigne de la co-indiciation)



Graphe de dépendance pour "Jean  
espere dormir"

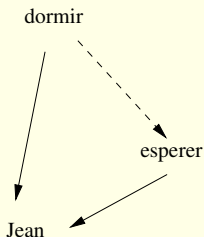
- ▶ Question : comment relier une information disponible dans l'arbre dérivé à un nœud de l'arbre de dérivation ?

⇒ *Intuition : Utiliser une structure combinant les deux structures de représentation des TAG : la Forêt Partagée*

⇒ *via une nouvelle opération : La fusion argumentale*

## La structure souhaitée : un graphe de dépendance ?

Le lien manquant est visible dans l'arbre dérivé (accord par percolation de traits témoigne de la co-indiciation)



Graphe de dépendance pour "Jean  
espere dormir"

- ▶ Question : comment relier une information disponible dans l'arbre dérivé à un nœud de l'arbre de dérivation ?

⇒ *Intuition : Utiliser une structure combinant les deux structures de représentation des TAG : la Forêt Partagée*

⇒ *via une nouvelle opération : La fusion argumentale*

## Le canevas de contrôle : Postulat

► **Postulat de départ :**

un arbre à contrôle transfère l'un de ses arguments vers l'arbre sur lequel il s'adjoit, celui-ci consomme cet argument.

Jean espère dormir

sujet d'espérer = sujet dormir

jean interdit à Marie de dormir

à-objet d'interdire = sujet de dormir

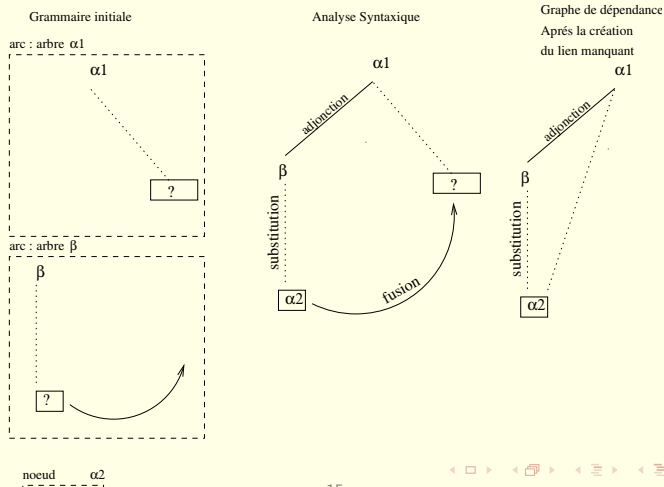
- L'information définissant quel argument est transféré est une information lexicale, disponible à travers un **canevas de contrôle** associé à une ancre lexicale.
- Lors de l'adjonction d'un verbe à contrôle, cet argument devra être fourni comme sujet de l'infinitive

## Le canevas de contrôle : Définition

- ▶ tout arbre auxiliaire  $\beta_c$  ancré par un verbe à contrôle possède une information lexicale qui indique quel argument, marqué par un indice  $i$  sur la catégorie d'un nœud  $X$  de cet arbre, doit être transféré vers un argument, marqué par un indice  $j$  sur le nœud de catégorie  $X$  d'un arbre
- ▶ on note ce transfert  $X_{i \rightarrow j}$

## Vision abstraite de l'opération de Fusion

Afin de construire un graphe, les nœuds du DT doivent être perçus comme des segments de graphes : un ensemble de nœuds et d'arcs



## Structure de la forêt partagée

- ▶ Une forêt partagée est l'intersection d'une grammaire et d'un automate d'entrée [Lang & Billot,89]
- ⇒ *Cette intersection est décrite avec une CFG (dans ce cas, la forêt ne génère correctement que les dérivations [Weir & Vijay-Shanker 93])*
- ▶ C'est aussi l'ensemble des items inférés lors d'une analyse tabulaire
- ⇒ *Notre forêt est composée de règles où chaque partie est un item CKY-like*
- ▶ Construction en 2 étapes : 1) On génère la forêt en tant qu'ensemble de règles 2) On exécute ces règles afin d'extraire toutes les dérivations



## Item de dérivations : vision globale

- *Chaque dérivation (substitution et adjonction) est marquée par des items définis par*  
 $\langle \text{Nœud, Arbre-Source, Arbres-Dest, Type de la dérivation} \rangle$
- ▶ **substitution** d'un arbre  $\alpha$  sur un nœud  $N$ , dont le canevas de contrôle est  $i \rightarrow j$ , d'un arbre  $\gamma$  :

$$D1 : \langle N_{i \rightarrow j}, \alpha, \gamma, \text{subs} \rangle$$

- ▶ **adjonction** d'un arbre  $\beta$  sur un nœud  $N$  d'un arbre  $\gamma$  :

$$D3 : \langle N, \beta, \gamma, \text{adj} \rangle$$

- ▶ **substitution** d'un arbre  $\alpha$  sur un nœud  $N$  d'un arbre  $\gamma$  :

$$D1' : \langle N, \alpha, \gamma, \text{subs} \rangle$$

- ▶ **dérivation axiomatique** d'un arbre initial  $\alpha$  de racine  $N$  :

$$D0 : \langle N, \alpha, -, - \rangle$$

## Une nouvelle dérivation : la substitution non réalisée

- *Une substitution non réalisée témoigne qu'un nœud  $N$  d'un arbre  $\gamma$  domine une chaîne vide*

▶ Substitution non réalisée :

$$D2 : \langle N, \boxed{?}, \gamma, subst \rangle$$

⇒ *L'ensemble de ces items est la forêt de dérivation*

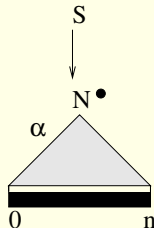
⇒ *On les infère quand chaque règle qui témoigne d'une dérivation est validée*

## Règles d'inférence (1)

Principe : Lors qu'on valide une règle témoignant d'une dérivation on infère un item de dérivation

### Dérivation de l'axiome

Si la règle  $r = S \longrightarrow \langle \top, N_\alpha, 0, n, -, -, \emptyset, \emptyset \rangle$  est validée lors de l'exécution de la grammaire, alors nous produisons l'item  $\langle N, \alpha, -, - \rangle$ .



$$\frac{S \longrightarrow \langle \top, N_\alpha, 0, n, -, -, \emptyset, \emptyset \rangle}{\langle N, \alpha, -, - \rangle}$$

avec  $n = \text{longueur de la chaîne}$

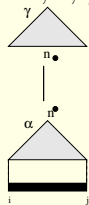
## Détail de l'extraction des dérivations : Règles d'inférence (2)

### Dérivation d'une substitution

Si la règle  $r = \langle \perp, n_\gamma, i, j, -, -, R, Pile \rangle \longrightarrow$

$\langle \top, n_\alpha, i, j, -, -, R, Pile \rangle$  est validée alors nous produisons l'item

$\langle N, \alpha, \gamma, subst \rangle$ .

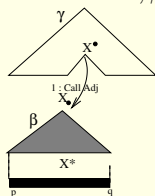


$$\frac{\langle \perp, n_\gamma, i, j, -, -, R, Pile \rangle \longrightarrow \langle \top, n_\alpha, i, j, -, -, R, Pile \rangle}{\langle N, \alpha, \gamma, subst \rangle}$$

## Règles d'inférence (3)

### Dérivation d'une adjonction

Si la règle  $r = \langle \top, n_\gamma, i, j, -, -, R, Pile \rangle \longrightarrow \langle \top, n_\beta, i, j, -, -, R, Pile' \rangle$  est validée alors nous produisons l'item  $\langle N, \beta, \gamma, adj \rangle$ .

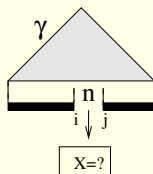


$$\frac{\langle \top, n_\gamma, i, j, -, -, R, Pile \rangle \longrightarrow \langle \top, n_\beta, i, j, -, -, R, Pile' \rangle}{\langle N, \beta, \gamma, adj \rangle}$$

## Règles d'inférence (4)

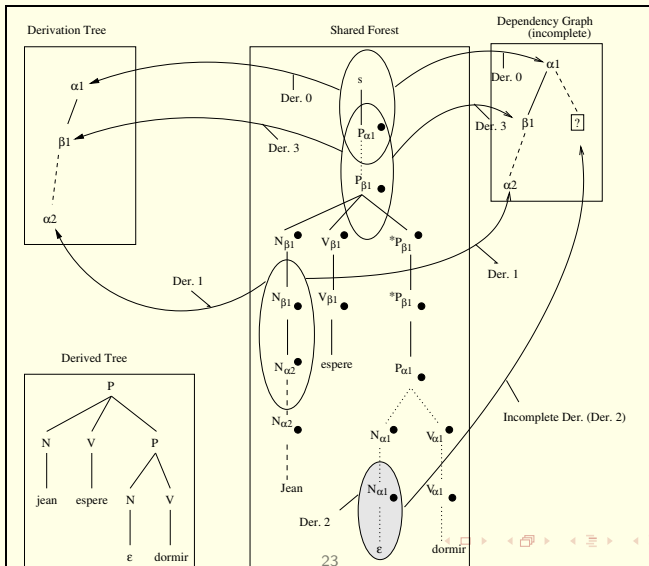
### Dérivation d'une substitution non réalisée

Si la règle  $\langle \perp, N_\gamma, i, j, -, -, R, Pile \rangle \rightarrow \text{vrai}$  est validée  
alors on produit l'item  $\langle N, \alpha, \boxed{?}, subst \rangle$ .



$$\frac{\langle \perp, N_\gamma, i, j, -, -, R, Pile \rangle}{\langle N, \alpha, X, subst \rangle} \text{ avec } X, = \boxed{?}$$

# Vision globale de la forêt partagée



## Création du lien manquant via l'opération de fusion

- ▶ Soit D1, l'item  $\langle N, \beta, \gamma, adj \rangle$  preuve de l'adjonction de  $\beta$  sur le nœud  $N$  de  $\gamma$  et  $C$  the canevas de contrôle  $\boxed{i \rightarrow j}$  de l'arbre  $\beta$
- ▶ D2, l'item  $\langle N_i, \alpha, \beta, subst \rangle$  preuve de la substitution de  $\alpha$  sur le nœud de substitution  $i$  de  $\beta$
- ▶ D3, l'item  $\langle N_j, \boxed{?}, \gamma, subst \rangle$  preuve d'une substitution non réalisée sur le nœud  $j$  sur l'arbre  $\gamma$  où  $\beta$  s'est adjoint

La Fusion argumentale est donc la règle d'inférence :

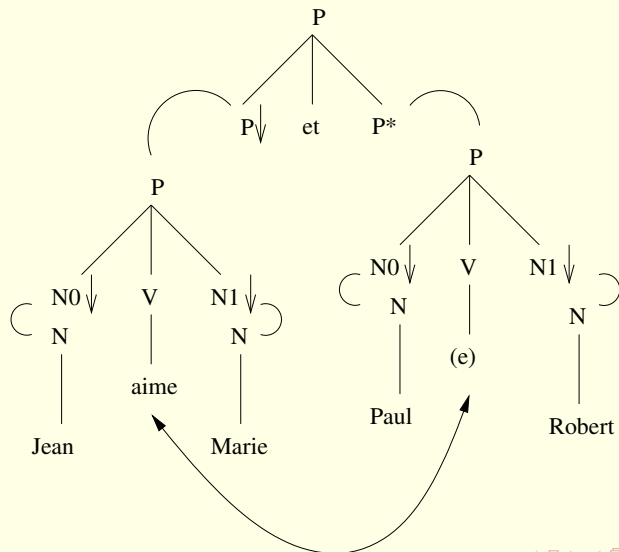
$$\frac{\langle N, \beta, \gamma, adj \rangle \langle N_{i \rightarrow j}, \alpha, \beta, subst \rangle \langle N_j, \boxed{?}, \gamma, subst \rangle}{\langle N_j, \alpha, \gamma, subst \rangle}$$



## Conclusion

- ▶ L'ensemble final des items contient les liens de coindexations
- ▶ Point de départ de post-traitements (liens inversés, adjonction multiple sur un nœud) pour finaliser le graphe de dépendance à *la* (Rambow & Joshi (94))
- ▶ La forêt décrit toutes les analyses et donc est une Forêt de dépendance

## Coordination élliptique (gap dependency)



1. pas d'information lexicale
2. Un ancrage non réalisé mais lien présent (trace d'accord et sous-cat)
3. ici on transfère une dérivation **implicite : l'ancrage**
4. **Meme processus que la fusion sauf qu'il ne s'agit plus d'une variable à instancier**
5. **mais d'un arbre y compris dans le transfert des structures associées**