Marcel Cori
Université Paris X - Nanterre

Arbres
et discontinuité

les 10 et 17 décembre 2004

La suite :

Définir précisément un objet privilégié de la formalisation en linguistique : l ’arbre.

Qu’est-ce qu’un arbre ? y a-t-il différentes sortes d ’arbres? Quelles sont les limites des arbres dans la représentation en syntaxe ?

A travers l’étude d’un problème : la discontinuité:

Paul a, le pauvre, Marie en a pleuré, perdu son emploi.

1. Introduction

2. Les arbres: définitions

3. La discontinuité et ses représentations

1. Introduction

On a l ’habitude de représenter les données linguistiques par des arbres. Mais:

(1) On ne sait pas vraiment ce que sont les arbres:
- il y a plusieurs sortes d ’arbres;
- il y a plusieurs définitions possibles des (mêmes) arbres;
- il y a une différence entre la structure effective qu’on représente, et le moyen de la représenter.

(2) Le phénomène de la discontinuité remet en cause l’usage des arbres et oblige à être très précis sur les outils utilisés.

Les différentes sortes d'arbres

(1) Arbres syntagmatiques

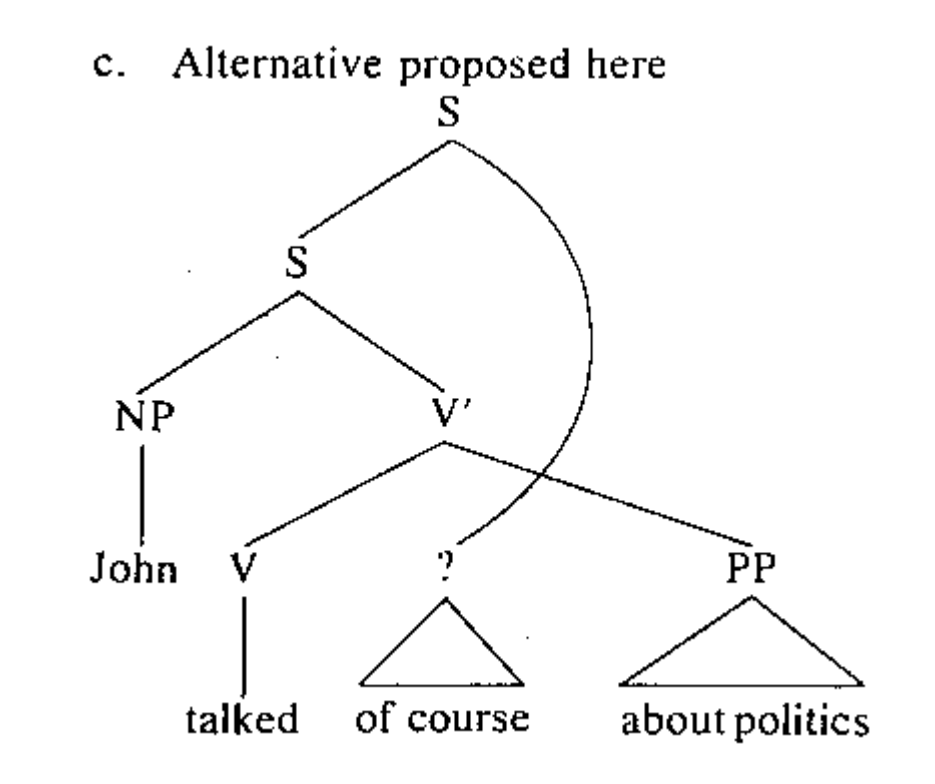
(2) Arbres de dépendances

(3) Arbres de représentation des connaissances encyclopédiques

(4) Arbres de structuration d 'un document

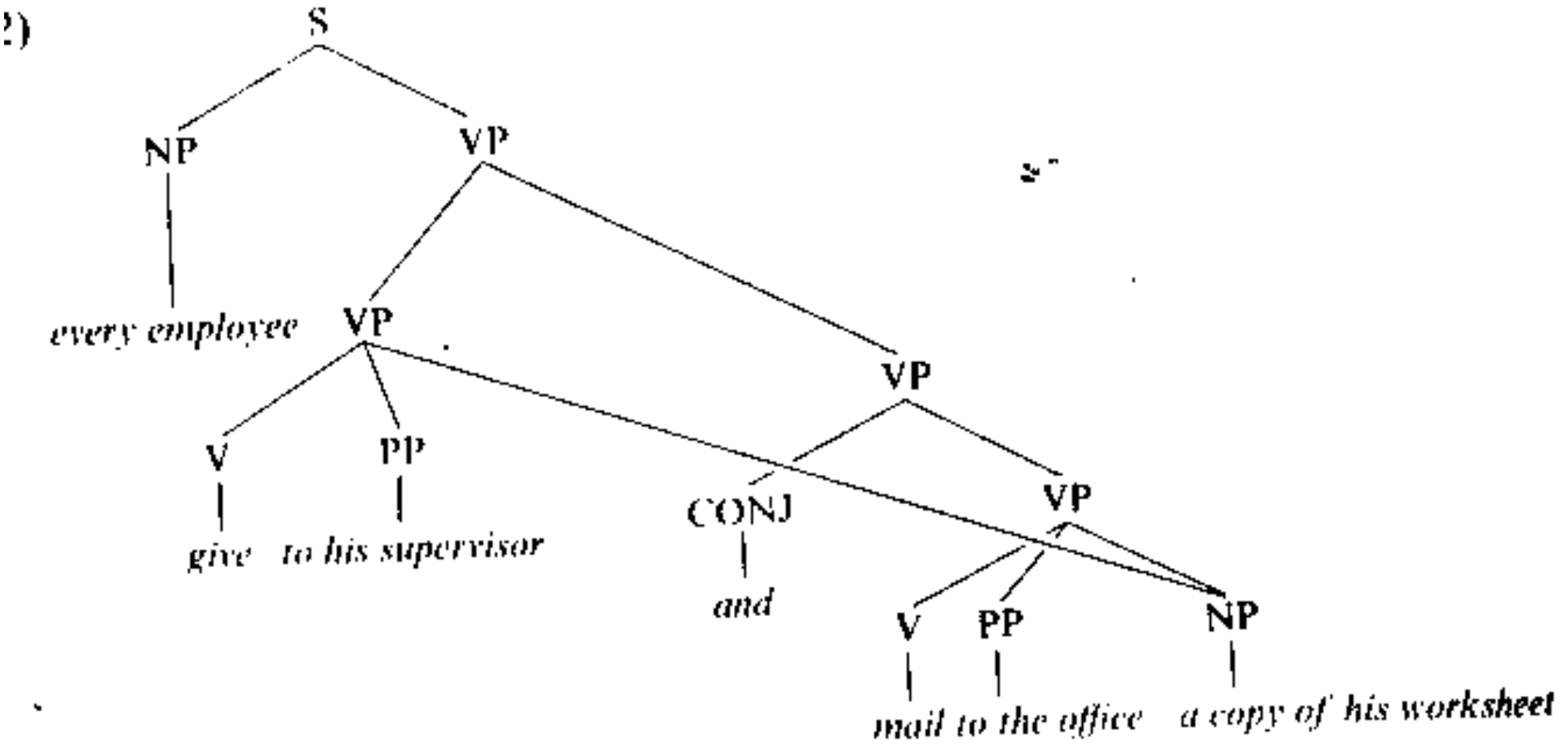
(5) Arbre généalogique

La discontinuité



McCawley, 1982

Ojeda, 1987


S
NP
VP
every employee
VP
VP
V
PP
CONJ
VP
give
to his supervisor
and
V
PP
NP
mail to the office
a copy of his worksheet

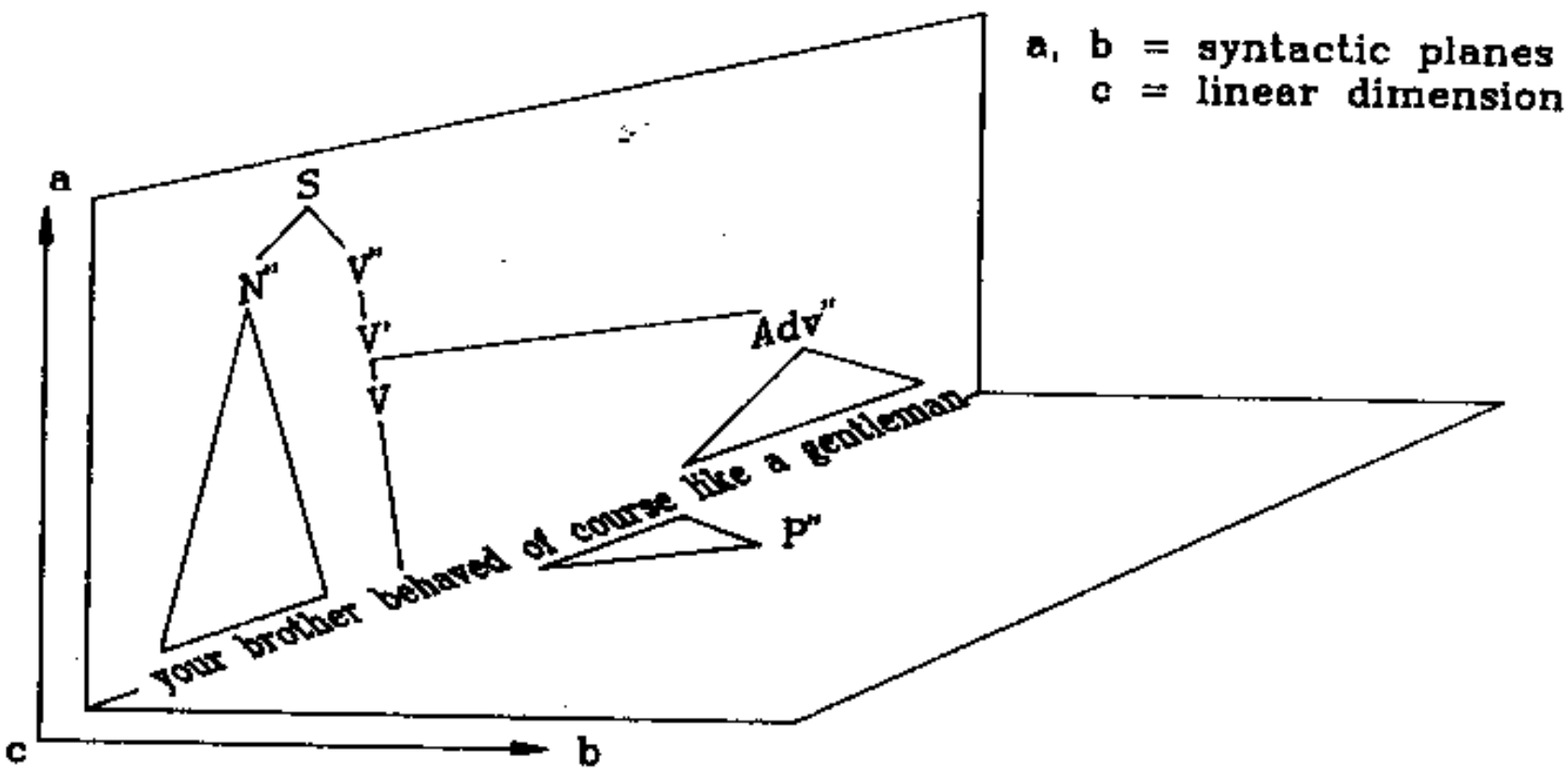


FIGURE 5.

Espinal, 1991

2. Les arbres : définitions

2.1 Définition à base de graphes

2.2 Définition alternative

2.3 Arbres et grammaires

2.4 Les représentations en machine des arbres

2.5 Arbres et structures parenthésées

3. La discontinuité et ses représentations

3.1. Le problème linguistique

3.1.1 Les données linguistiques

3.1.2 Caractérisation de l'incidence

3.2 Les strctures représentatives de la discontinuité

3.3 Une formalisation de ces structures

3.1. Le problème

3.1.1 Les données linguistiques

a. *He waked your friend up.*

b. *John talked, of course, about politics.*

c. *Jean a vraisemblablement oublié de faire son devoir.*

d. *Pierre a, le pauvre, perdu son emploi.*

e. *Pierre a, le pauvre, Marie en a pleuré, perdu son emploi.*

f. *Pierre a, Marie la pauvre en a pleuré, perdu son emploi.*

Discontinuité: la suite de mots constituant un syntagme est coupée en deux parties par un ou plusieurs autres syntagmes.

He waked your friend up.

Le constituant *waked up* ne peut se passer du constituant *your friend*.

Autres cas : on conserve une phrase bien formée lorsqu'on supprime le *constituant incident* (désormais CI).

Pierre a, le pauvre, perdu son emploi.

le pauvre est le constituant incident (CI).

On appellera *incidence* le phénomène. La structure (interrompue) dans laquelle est inséré le CI sera la *structure hôte* :

Pierre a perdu son emploi.

La position où le CI est inséré sera appelée le *site d'incidence* (SI):

*Pierre a * perdu son emploi.*

Il y a des cas particuliers d'incidence sans discontinuité :

Pierre est intelligent, je trouve.

Le pauvre, Pierre a perdu son emploi.

⇒ L'incidence et la discontinuité sont deux phénomènes distincts, qui ne se recouvrent que partiellement.

Dans la suite, on s'intéresse à l'incidence, en supposant que d'autres cas de discontinuité pourraient être représentés à l'aide des mêmes outils formels.

3.1.2 Caractérisation de l'incidence

Cf. (Espinal, 1991), (Cori et Marandin, 1995).

- (A) Les CI appartiennent à toutes les catégories majeures : SN, S, S', SA, SAdv, SP.

Jean a vraisemblablement oublié de faire son devoir.
Pierre a, le pauvre, perdu son emploi.
Pierre a, Marie en a pleuré, perdu son emploi.
Pierre a, en six mois, perdu trois kilos.

- (B) Les seules parties de phrase où l'incidence est impossible en français est entre l'article et le nom dans le SN, entre le clitique et le verbe dans le SV, entre la préposition (*à, de*) et le SN dans le SP.

* *Il vraisemblablement a oublié de faire son devoir.*
* *Le, Marie le trouve très énervant, professeur a fait son cours.*

- (C) Les CI n'entrent pas dans des relations grammaticales avec les autres constituants dans la phrase hôte.

(D) L'incidence est soumise à des contraintes d'occupabilité :

Pierre a, le pauvre, perdu son emploi.
** Pierre a, le vieux, perdu son emploi.*

** Pierre n'a pas, bêtement, fait son devoir.*
? Pierre n'a pas, le fou, fait son devoir.
Pierre, bêtement, n'a pas fait son devoir.
? Pierre n'a pas fait son devoir, bêtement.
Pierre n'a pas fait son devoir, le fou.

- (E) Plusieurs constituants distincts peuvent apparaître dans un même site d'incidence :

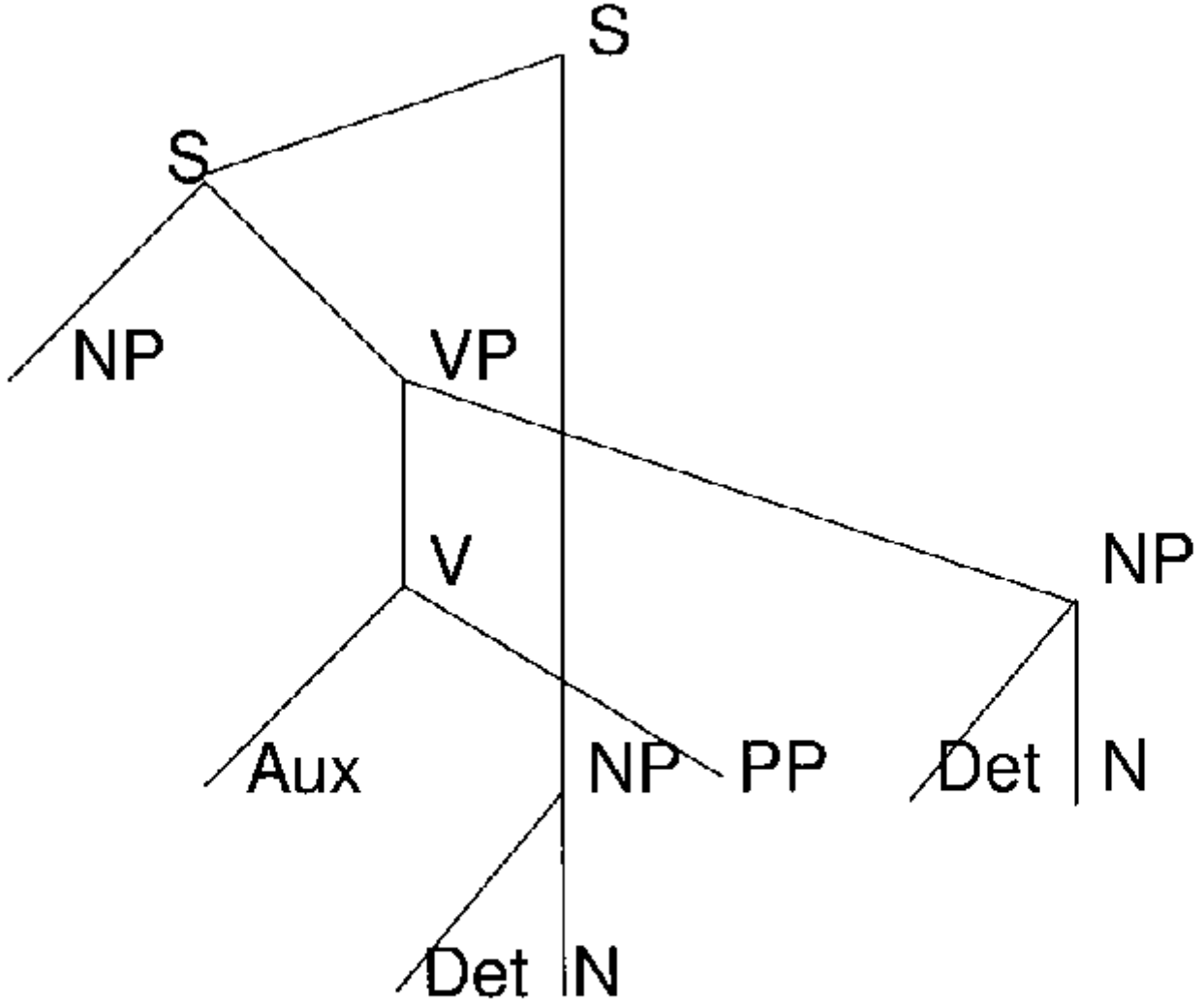
Pierre a, le pauvre, Marie en est toute retournée, perdu son emploi.

3.2 Les outils formels pour représenter la discontinuité

Les travaux les plus fréquents proposent une représentation de la discontinuité en général.

⇒ construction d'*arbres discontinus:* McCawley (1982), Bunt (1996).

Ces structures diffèrent des arbres syntagmatiques en ce que la condition de non-croisement et la condition d'exclusivité ne sont plus vérifiées.

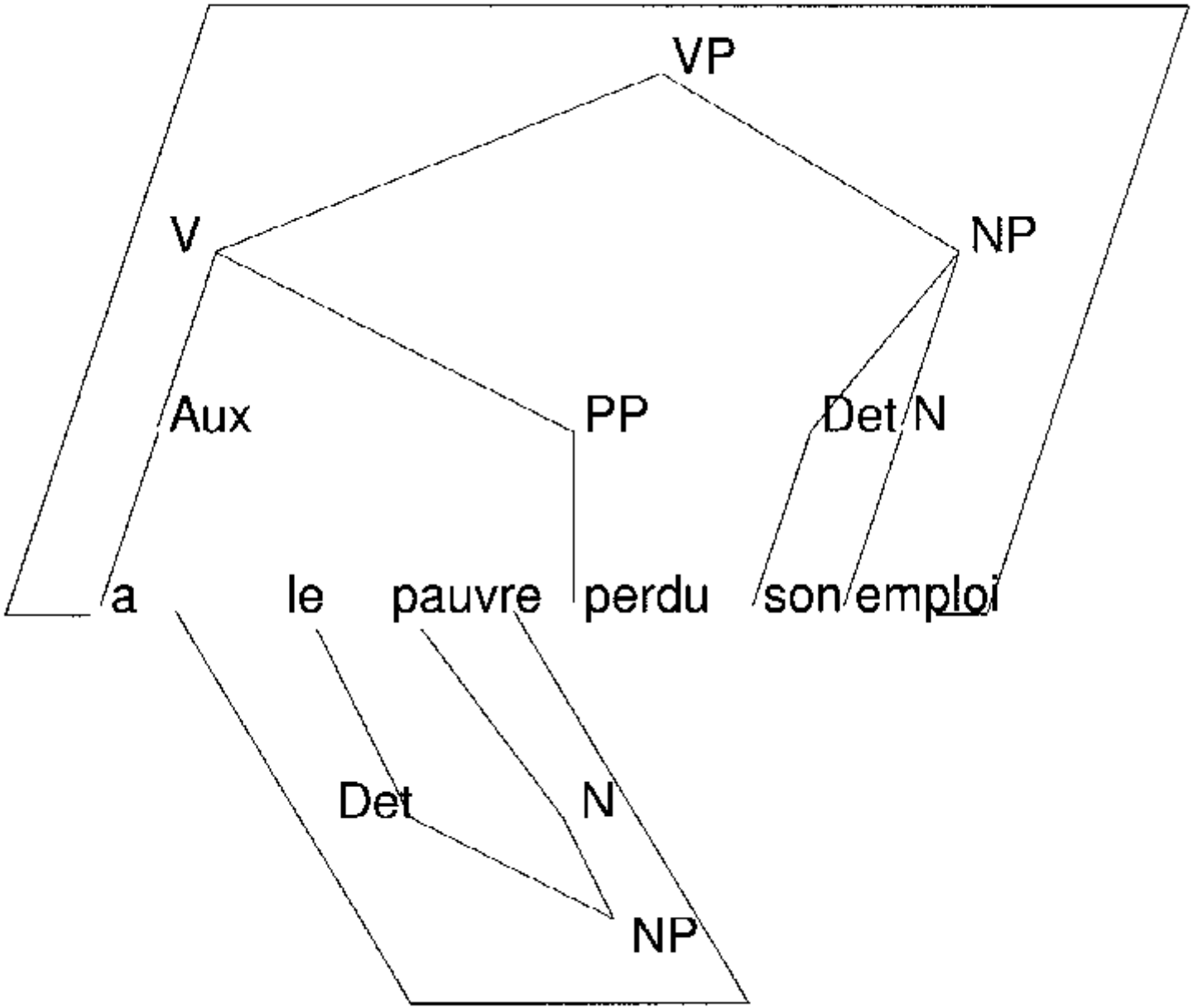


Pierre a le pauvre perdu son emploi

Thèse alternative : la structure hôte et le CI sont des structures syntaxiques indépendantes ; elles ne forment une unité qu'à un certain niveau de représentation sémantique.

Espinal (1991) : structure tridimensionnelle, STD. La structure hôte et le CI appartiennent à différents plans, qui ont leur intersection sur « la ligne où les relations structurales de précédence entre les différents sommets est spécifiée ».

⇒ une structure formée de plusieurs arbres distincts, dans laquelle l'ensemble des feuilles constitue un ensemble totalement ordonné.


VP
V
NP
Aux
PP
Det
N
a
le
pauvre
perdu
son
emploi
Det
N
NP

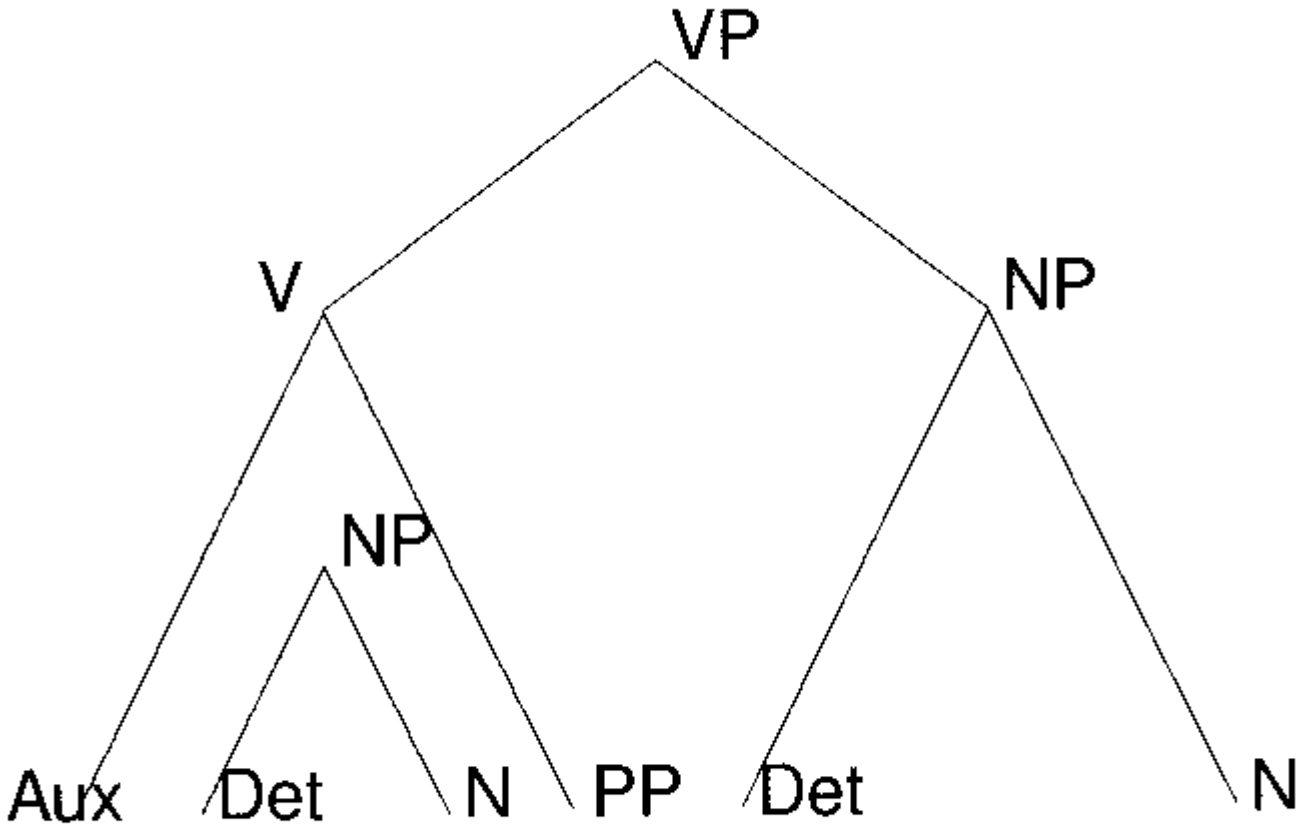
Les racines de ces arbres sont étiquetées par n'importe quelle catégorie majeure.

Deux CI, ou un CI et une structure hôte ne peuvent se chevaucher.

Nous considérerons qu'un CI peut être la structure hôte d'un autre CI, et ainsi de suite :

Jean a, Paul tous les jours me le répète, tort de fumer.

3.3 Une formalisation des STD



3.3.1 Définition des STD

Cat est un ensemble fini d'étiquettes. Une STD est définie comme étant un quadruplet $A = \langle X, L, D, P \rangle$ où :

(a) X est un ensemble fini, l'ensemble des sommets;

(b) $L : X \rightarrow Cat$ est la fonction d'étiquetage ;

(c) $D \subseteq X \times X$ est un ordre partiel, la relation de dominance ;

(d) $P \subseteq X \times X$ est un ordre partiel, la relation de précédence,

et tel que quatre conditions soient vérifiées.

Définitions (incidentes) :

Ensemble des feuilles de l'arbre :

$$feuilles(A) = \{x \in X;\ ^{\forall} y \in X\ \langle x,y\rangle \in D \Rightarrow x=y\}$$

Ensemble des feuilles dominées par un sommet :

$$feuilles(x) = \{y \in feuilles(A)\ ;\ \langle x,y\rangle \in D\}$$

Conditions:

→ L'exclusivité faible (McCawley) remplace l'exclusivité

(i) $\forall\, x,y \in X\ (\langle x,y\rangle \in P \vee \langle y,x\rangle \in P)$
$\Rightarrow (\langle x,y\rangle \notin D \wedge \langle y,x\rangle \notin D)$

→ Il y a un ordre total sur l'ensemble des feuilles :

(ii) $\forall\, x,y \in \mathit{feuilles}(A)\ \ x \neq y \Rightarrow (\langle x,y\rangle \in P \vee \langle y,x\rangle \in P)$

→ un sommet x en précède un autre y si et seulement si toutes les feuilles dominées par x précèdent toutes les feuilles dominées par y :

(iii) $\forall\, x,y \in X\ \langle x,y\rangle \in P \Leftrightarrow$
$(\forall\, u \in \mathit{feuilles}(x)\ \forall\, v \in \mathit{feuilles}\,(y)\ \langle u,v\rangle \in P)$

→ Conditions pour qu'un constituant puisse être inséré dans une structure hôte :

(iv) Pour tous les $x, y \in X$ tels que $\langle x,y\rangle \notin D$ et $\langle y,x\rangle \notin D$ et $\langle x,y\rangle \notin P$ et $\langle y,x\rangle \notin P$,

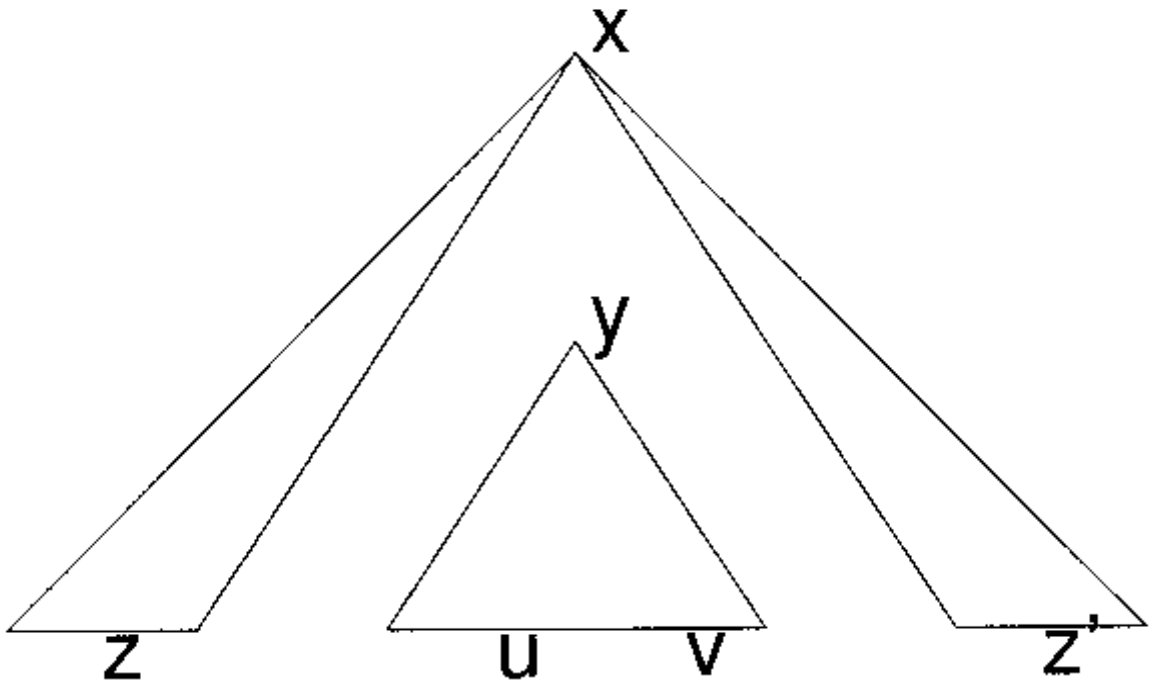
une des deux conditions suivantes doit être satisfaite:

(iv.1) $\forall\, z \in feuilles(x)\ \ \forall\, u,v \in$ feuilles(y)

$(\langle z,u\rangle \in P \wedge \langle z,v\rangle \in P) \vee (\langle u,z\rangle \in P \wedge \langle v,z\rangle \in P)$

(iv.2) $\forall\, z \in feuilles(y)\ \ \forall\, u,v \in feuilles(x)$

$(\langle z,u\rangle \in P \wedge \langle z,v\rangle \in P) \vee (\langle u,z\rangle \in P \wedge \langle v,z\rangle \in P)$



3.3.2 Propriétés

Propriété 1 La condition de non-croisement est vérifiée :

$$\forall\, w,x,y,z \in X$$
$$\langle w,x\rangle \in P \wedge \langle w,y\rangle \in D \wedge \langle x,z\rangle \in D \Rightarrow \langle y,z\rangle \in P$$

Propriété 2 La condition de la mère unique est vérifiée:

$$\forall\, x,y,z \in X$$
$$\langle x,z\rangle \in D \wedge \langle y,z\rangle \in D \Rightarrow (\langle x,y\rangle \in D \vee \langle y,x\rangle \in D)$$

Définition $y \in \text{fils}(x)$ si et seulement si $\langle x,y\rangle \in D$ et

$$\forall z \in X \ \langle x,z\rangle \in D \wedge \langle z,y\rangle \in D \Rightarrow (x = z \vee y = z)$$

Propriété 3 Il y a un ordre total sur les fils d'un même sommet :

$$\forall x \in X \ \forall y,z \in \text{fils}(x) \ (\langle y,z\rangle \in P \vee \langle z,y\rangle \in P)$$

La définition de l'admissibilité des arbres est fondée sur cette dernière propriété.

3.3.3 Grammaires et admissibilité

Une grammaire est donnée par $G = \langle Cat, Cat_M, R, R_I \rangle$, où :

- Cat_M est le sous-ensemble de *Cat* formé des catégories majeures ;

- R contient des règles telles que $\alpha \rightarrow \beta\ \gamma$ avec $\alpha, \beta, \gamma \in$ *Cat* : la grammaire est supposée sous forme normale de Chomsky ;

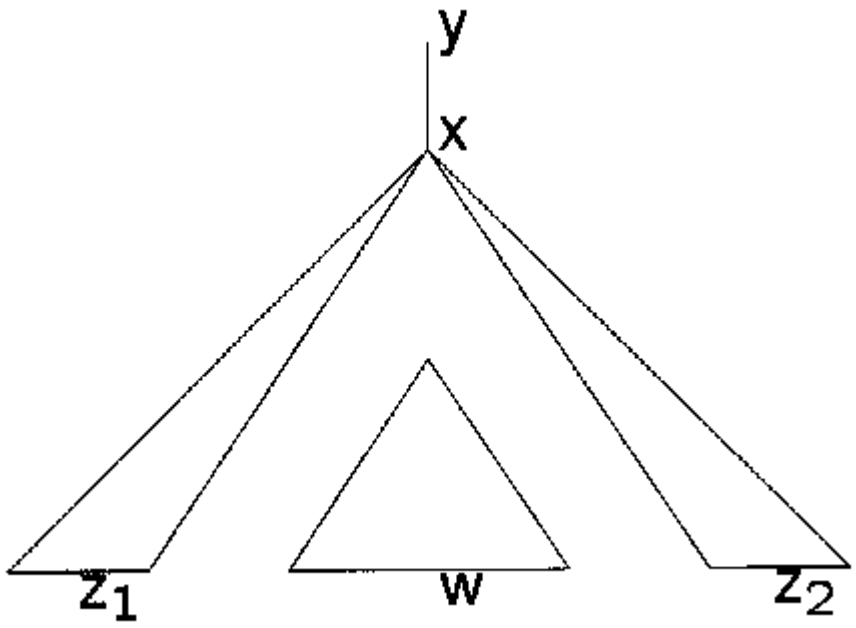
- R_I est un sous-ensemble de R : les règles autorisées à recevoir un CI.

Définition: Un sommet x (representant un syntagme) d'une STD *A* est un *sommet hôte* si et seulement si il existe deux sommets $z_1, z_2 \in$ *feuilles*(*x*) tels que :

(i) $\exists\, w \in$ *feuilles*(*A*) - *feuilles*(x)

$\langle z_1, w\rangle \in P \wedge \langle w, z_2\rangle \in P$

(ii) $\forall\, y \;\; z_1, z_2 \in$ *feuilles*(y) $\Rightarrow \langle y, x\rangle \in D$



Définition Une STD est admissible par la grammaire si et seulement si

(1) chacun de ses sous-arbres élémentaires est compatible avec une des règles de R, et

(2) la règle est dans R_I quand la racine du sous-arbre élémentaire correspondant est un sommet hôte.

3.3.4 Généralisation des définitions

Cat est un ensemble fini d'étiquettes. Une structure discontinue (SD) est définie comme étant un quadruplet $A = \langle X, L, D, P \rangle$ où :

(a) X est un ensemble fini, l'ensemble des sommets;

(b) $L : X \rightarrow Cat$ est la fonction d'étiquetage ;

(c) $d \subseteq X \times X$ est une relation anti-réflexive, la relation de dominance directe ;

(d) $P \subseteq X \times X$ est un ordre partiel, la relation de précédence,

et tel que quatre conditions soient vérifiées.

Conditions:

→ La fermeture transitive et réflexive D de d est une relation d 'ordre partiel

(1) $\forall\, x,y \in X\ (\langle x,y\rangle \in D \wedge \langle y,x\rangle \in D) \Rightarrow x = y$

→ Exclusivité faible :

(2) $\forall\, x,y \in X\ (\langle x,y\rangle \in P \vee \langle y,x\rangle \in P)$

$\Rightarrow (\langle x,y\rangle \notin D \wedge \langle y,x\rangle \notin D)$

→ Il y a un ordre total sur l'ensemble des feuilles :

(3) $\forall\ x,y \in \mathit{feuilles}(A)\ \ x \neq y$
$\Rightarrow (\langle x,y\rangle \in P \vee \langle y,x\rangle \in P)$

→ un sommet x en précède un autre y si et seulement si toutes les feuilles dominées par x précèdent toutes les feuilles dominées par y :

(4) $\forall\ x,y \in X\ \langle x,y\rangle \in P \Leftrightarrow$
$(\forall\ u \in \mathit{feuilles}(x)\ \ \forall\ v \in \mathit{feuilles}(y)\ \langle u,v\rangle \in P)$

Conditions supplémentaires:

→ Condition de la mère unique :

(5) $\forall\, x,y,z \in X\ (\langle x,z\rangle \in d \wedge \langle y,z\rangle \in d) \Rightarrow x = y$

→ Existence d 'une racine (unique) :

(6) $\exists\, x \in X \quad \forall\, y \in X\ \langle x,y\rangle \in D$

→ Exclusivité (non faible) :

(7) $\forall\, x,y \in X\ (\langle x,y\rangle \in P \vee \langle y,x\rangle \in P)$
$\Leftrightarrow (\langle x,y\rangle \notin D \wedge \langle y,x\rangle \notin D)$

→ Condition de non-croisement :

(8) $\forall\, w,x,y,z \in X$
$\langle w,x\rangle \in P \wedge \langle w,y\rangle \in D \wedge \langle x,z\rangle \in D \Rightarrow \langle y,z\rangle \in P$

→ Condition de non-chevauchement :

(9) Pour tous les $x, y \in X$ tels que $\langle x,y\rangle \notin D$ et $\langle y,x\rangle \notin D$ et $\langle x,y\rangle \notin P$ et $\langle y,x\rangle \notin P$,
une des deux conditions suivantes doit être satisfaite :

(9.1) $\forall\ z \in \mathit{feuilles}(x)\ \ \forall\ u,v \in \mathrm{feuilles}(y)$
$(\langle z,u\rangle \in P \wedge \langle z,v\rangle \in P) \vee (\langle u,z\rangle \in P \wedge \langle v,z\rangle \in P)$

(9.2) $\forall\ z \in \mathit{feuilles}(y)\ \ \forall\ u,v \in \mathit{feuilles}(x)$
$(\langle z,u\rangle \in P \wedge \langle z,v\rangle \in P) \vee (\langle u,z\rangle \in P \wedge \langle v,z\rangle \in P)$

Propriétés

Propriété 1 (4) $\Rightarrow$ (8)

Propriété 2 (7) $\Rightarrow$ (2)

Propriété 3 (7) $\Rightarrow$ (3)

Propriété 4 (7) et (8) $\Rightarrow$ (4)

Définition de différentes structures:

(a) les arbres discontinus (McCawley): (5) et (6)

(b) les arbres discontinus avec multidominance (Ojeda): (6)

(c) les structures tridimensionnelles : (5) et (9)

(d) les arbres (Wall) : (5), (6), (9)

Conclusion

Pour faire du TAL, on a besoin :

- d'analyses linguistiques fines,
- d'un formalisme logico-mathématique précis,
- de structures de données informatiques précisément définies,
- de programmes efficaces.