Marcel Cori
Université Paris X - Nanterre

Traitement automatique des langues

et formalisation en linguistique

le 29 octobre 2004

1. Du traitement automatique des langues à la formalisation en linguistique

2. Les ambiguïtés en TAL

3. Une forme spécifique d ’ambiguïté

1. Du traitement automatique des langues à la formalisation en linguistique

Définition : le ***traitement automatique des langues (TAL)*** est constitué des méthodes et des programmes qui prennent pour données des productions langagières, quand ces méthodes et programmes tiennent compte des spécificités des langues humaines.

Les oppositions internes au TAL

(1) L ’opposition objectifs pratiques / recherche théorique

objectifs pratiques = objectifs industriels

(2) L ’opposition TAL robuste / TAL théorique

(3) L ’opposition méthodes numériques / méthodes symboliques

TAL robuste / TAL théorique

TAL robuste

- s'applique à de vraies productions langagières
- doit donner des résultats quelles que soient les données

TAL théorique

- les données peuvent être construites par le linguiste
- objectif : fournir des systèmes de description des langues

Les perspectives du TAL théorique

En traitant des données idéales, tenter de séparer les différents phénomènes, les différentes propriétés :

- morphologie,
- syntaxe,
- sémantique, …

TAL théorique = TAL de laboratoire,

TAL expérimental

Informatisation d'un problème ⇒ nécessité d'être explicite, précis et objectif.

Les règles énoncées entrent dans des processus automatisés ⇒ impossibilité de rester dans le vague ou d'être ambigu.

D'où la construction de systèmes de description des langues rigoureux, ou le perfectionnement de systèmes existants.

Le TAL entraîne plus de rigueur dans la recherche en linguistique.

L ’opposition procédural/ déclaratif

Approche procédurale:

```
if a[0]=='V':
      if categ(a[1:])=='SN': return 'SV'
      elif categ(a[1:])=='SP': return 'SV'
```

Approche déclarative:

a. SV → V SN

b. SV → V SP

Les traitements déclaratifs

- Mise à jour facilitée:

c. SV → V SN SP

- division du travail entre l'informaticien (qui écrit les programmes) et le linguiste (qui définit les grammaires et les lexiques)

Logique des « systèmes experts ».

Nécessité de trouver un langage commun entre le linguiste et l ’informaticien :

c ’est le modèle formel dans lequel le linguiste va exprimer ses données,

et sur lequel vont agir les programmes de l ’informaticien.

Le modèle formel doit être:

- rigoureux : les objets doivent être bien définis ;
- expressif : il doit être lisible par un être humain.

⇒ La formalisation en linguistique

2. Les ambiguïtés en TAL

Dans l'analyse automatique d'un énoncé, la première tâche consiste en une *segmentation* de l 'énoncé en unités de base (mots ?), et en l '*étiquetage* de ces unités.

Problèmes : La polysémie, l'homophonie ou l'homographie.

Le boucher ferme son livre

Mon beau-frère est masseur
Mon beau-frère et ma sœur
Mon beau-frère est ma sœur
Mon beau-frère aima sœur

Beaucoup d ’ambiguïtés sont résolues par la syntaxe :

Le boucher ferme son livre
Mon beau-frère aima sœur
Mon beau-frère et ma sœur

Mais, il reste des ambiguïtés que la syntaxe ne peut résoudre :

Mon beau-frère est masseur
Mon beau-frère est ma sœur

Enfin, il y a des énoncés où l ’étiquetage est non ambigu, mais dont la syntaxe est ambiguë, et donc que l ’on peut comprendre de plusieurs façons.

Ambiguïtés syntaxiques

J’ai rencontré une directrice de société enrhumée
J’ai rencontré un conducteur de train électrique
J’ai trouvé une pièce de monnaie ancienne

Jean lit le livre
Jean lit le matin
Jean mange le matin
Jean aime le matin

“ Attachement prépositionnel ”

Il mange une glace au chocolat
Il mange une glace au restaurant
Il mange une glace au café

La statue de marbre de Rodin du duc d ’Aumale de retour d ’Afrique

D ’où le rejet par le TAL robuste de la résolution de certaines ambiguïtés.

Enlève le bonnet du bébé et mets-le à la machine à laver
Enlève le bonnet du bébé et mets-le au lit

Jean-Louis Debré : « Je préfère la constitution de mon père à celle de Mamère »

D’où l’intérêt de la sémantique, de la pragmatique, …

3. Une forme spécifique d ’ambiguïté

a. Les rayonnements magnétiques perturbent *les électriques*
b. Il a mangé *les pourries*
c. *Le parler vrai* du ministre lui a causé des ennuis

d. Il a un veston *très sport*
e. Paul est *très sieste*

f. *Que tu viennes* m'ennuie
g. *Le frapper* pourrait nous valoir des ennuis

Analyse non TAL :

Les rayonnements magnétiques perturbent les ∅ électriques
Il a mangé les ∅ pourries
(∅ de catégorie N)

∅ que tu viennes m'ennuie (∅ de catégorie SN)
∅ ∅ le frapper pourrait nous valoir des ennuis

Le parler vrai du ministre lui a causé des ennuis
(parler de catégorie N)

Il a un veston très sport
Paul est très sieste

- Les « objets vides » sont très difficiles à traiter en analyse syntaxique.

- Il n ’est pas raisonnable non plus de multiplier les catégories pour une même unité.

Syntagmes canoniques et syntagmes non canoniques

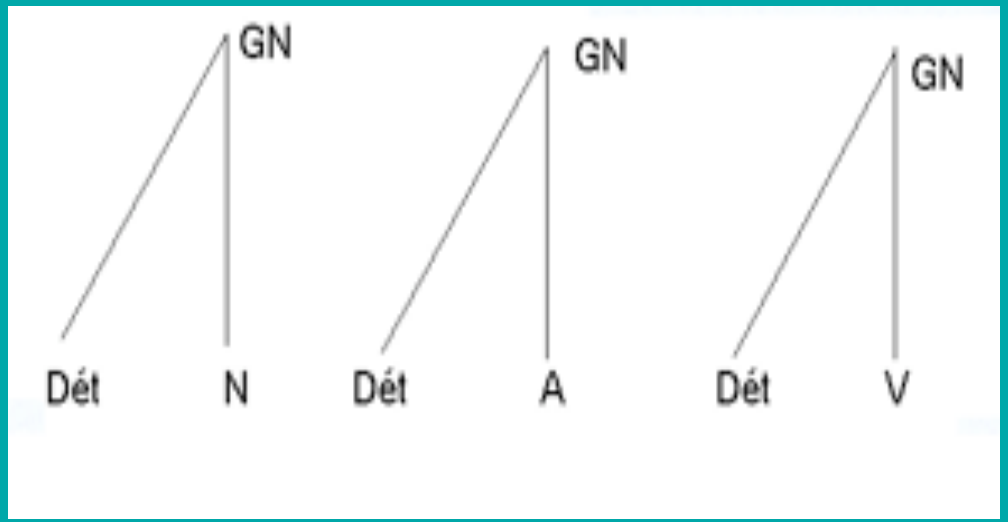
Un constituant de catégorie X apparaît quand on attend un constituant de catégorie Y.

Un syntagme avec un constituant non attendu sera dit **non canonique**.

Analyse du syntagme nominal

L'adjectif ou le verbe occupe la position noyau du GN.

- (i) il n'y a pas de différence structurelle entre un GN qui admet une tête nominale et un GN qui admet une tête verbale ou adjectivale : N, A ou V apparaissent exactement dans la même position.



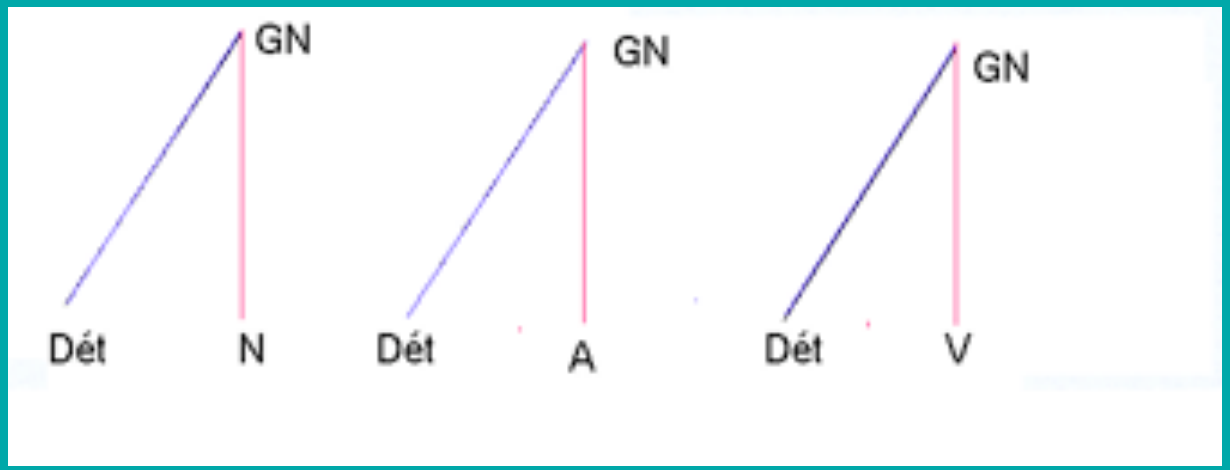
- (ii) le contraste entre les syntagmes canoniques et les syntagmes non canoniques n'est pas corrélé à une différence de structures, il est lié au fait qu'un N est plus naturel qu'un A ou un V dans la position noyau d'un GN.

Représentation

Les positions doivent être définies sans référence à la catégorie du constituant qui les occupe ; la position noyau ne fait pas exception.

C'est pourquoi on introduit des *couleurs* dans la définition des arbres syntaxiques.

Coloration des arbres :

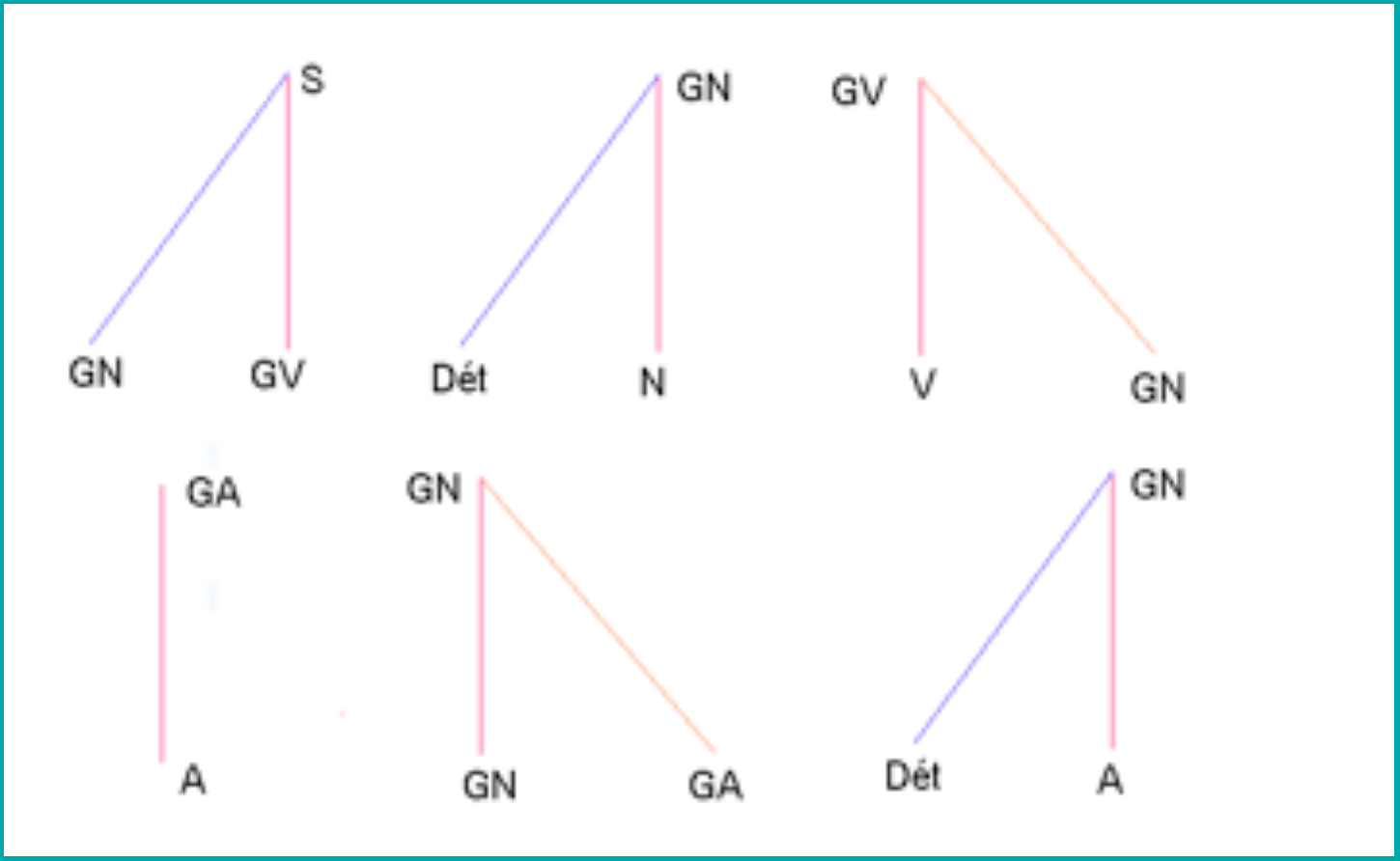


Une **grammaire d'arbres polychromes** est constituée par un ensemble fini d'arbres élémentaires à p couleurs:

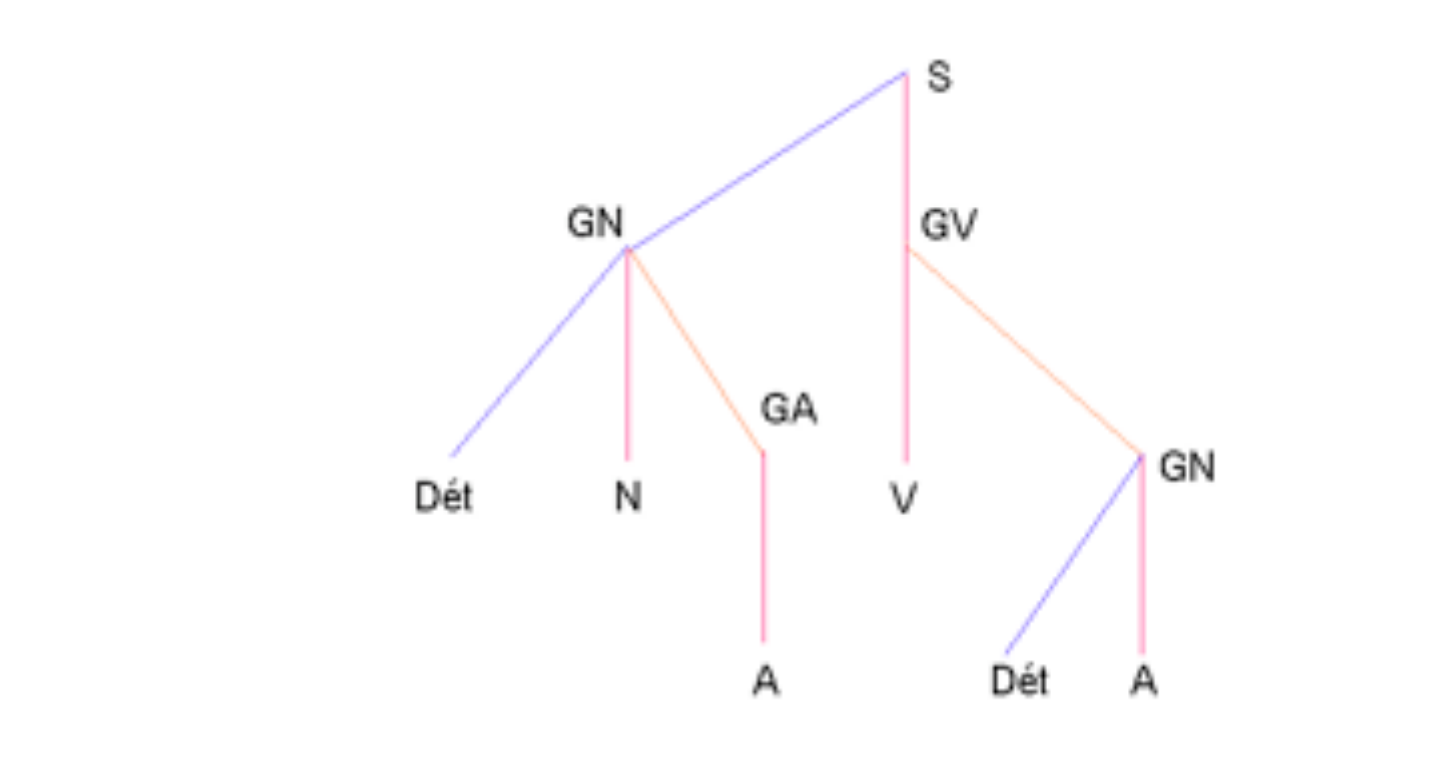
$$G = \{A_1, A_2, \ldots, A_m\}.$$

(Cori et Marandin, 1993, 1994, 1998)

Exemple de grammaire:
p = 5



Exemple d ’arbre polychrome *engendré* par la grammaire



Les rayonnements magnétiques perturbent les électriques

Les ambiguïtés induites

Marie a mangé *les mûres* (mûres: N ou A)
Marie a mangé les pourries
Marie a mangé les pommes

Paul est *très calme* (calme: A ou N)
Paul est très sieste
Paul est très intelligent

Le manger cru pourrait avoir des vertus thérapeutiques (le: pronom ou déterminant)

Le frapper pourrait avoir des vertus thérapeutiques
Le vin rouge pourrait avoir des vertus thérapeutiques
Les instinctivores préconisent le manger cru

Syntagmes préférés

Quand les énoncés sont interprétés isolément, l'interprétation fondée sur la structure canonique est préférée à l'interprétation fondée sur la structure non canonique.

Marie a mangé les mûres (mûres: N)

Paul est très calme (calme: A)

Le manger cru pourrait avoir des vertus thérapeutiques (le: déterminant)

Représentation

Le fait que certaines tournures soient plus naturelles que d'autres doit être pris en compte par la grammaire.

C'est pourquoi la grammaire est partitionnée en deux sous-ensembles d'arbres élémentaires disjoints:

$$G = C \cup N$$

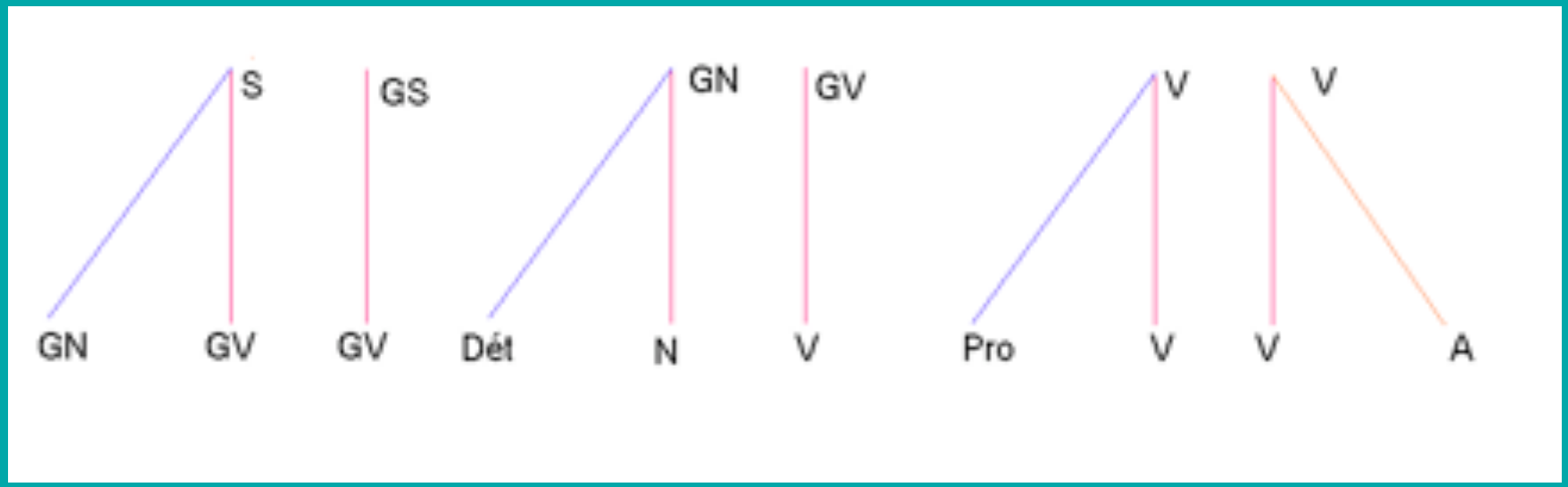
C est l'ensemble des structures **canoniques**

N est l'ensemble des structures **non canoniques**.

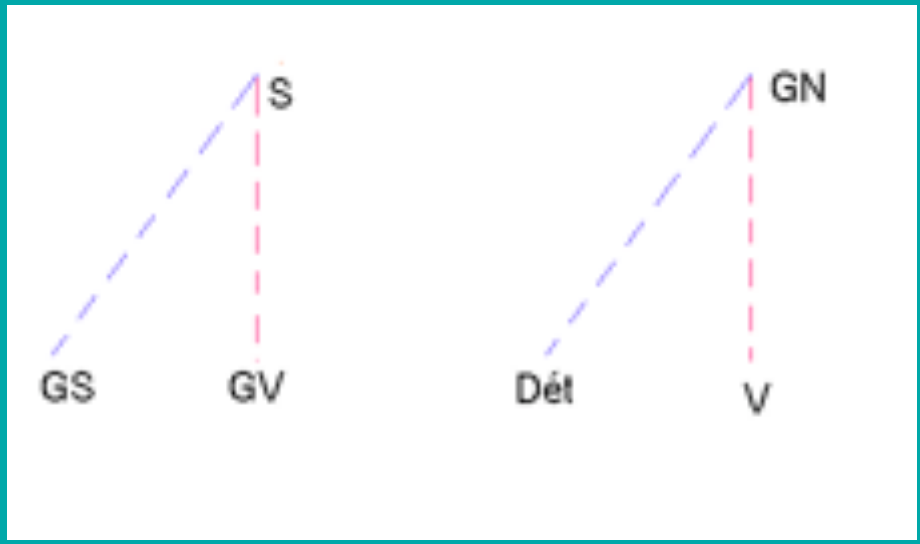
Ceci permet *un calcul de préférence.*

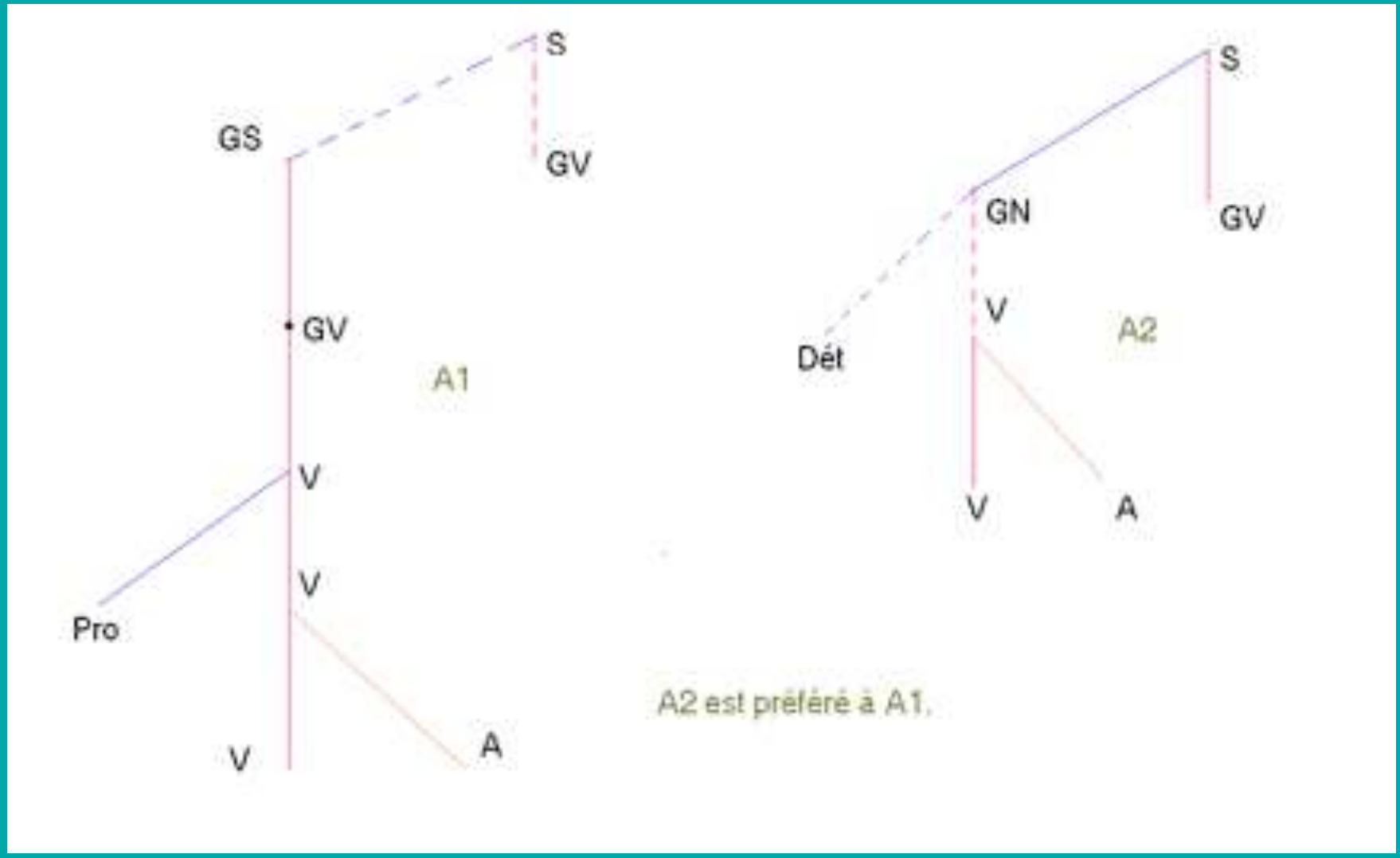
Cori et Marandin, 1997.

Arbres canoniques



Arbres non canoniques





Le manger cru aurait des vertus thérapeutiques

La préférence liée à la canonicité des structures enchâssantes l'emporte sur la préférence liée à la canonicité des structures enchâssées.

a. Pierre a apporté les pommes. Marie a mangé *les mûres*, Paul a jeté les vertes.

b. Il n'y a plus de feu pour cuire le bifteck. Le manger cru pourrait avoir des vertus thérapeutiques.

L'interprétation préférée est celle qui est corrélée avec une interprétation anaphorique.

Conclusion

La préférence syntaxique s'applique quand la lecture anaphorique ne convient pas.

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.