3 Annotation syntaxique

Questions:

1. Quelles infomations syntaxiques est-ce qu'on veut annoter?
2. Par quelle forme d'annotation est-ce qu'on peut exprimer ces informations syntaxiques?

(Plus tard, on va aussi parler des méthodes d'annotation syntaxique.)

Traits idéals d'un schema d'annotation:

- **Déscriptivité** : des phenomènes syntaxiques devraient être décrits plutôt que expliqués
- **Rapport aux données** (data-drivenness) : le schema d'annotation doit être assez riche pour décrire tous les phénomènes dans un corpus
- **Indépendance des théories** (theory neutrality) : le schema d'annotation ne doit pas être influencé par des considérations propres à une théorie spécifique. Par contre, il doit contenir assez d'information pour extraire des représentations conforme à une théorie.

3.1 Propriétés syntaxiques à annoter

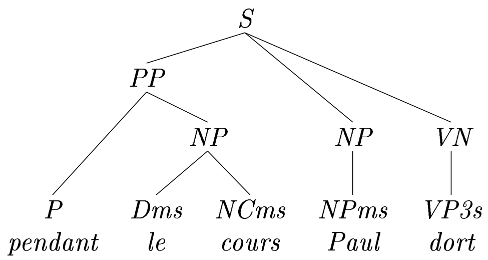
3.1.1 Constituants

Annotations des groupes de mots qui constituent des syntagmes.

Une structure de constituants peut être representée par un arbre.

Un arbre est un ensemble de noeuds avec des relations binaires de dominance immédiate et precédence linéaire immédiate. Les noeuds internes d'un arbre de constituance sont décorés avec des étiquettes (des catégories syntaxiques). Les feuilles sont décorées avec les mots de la phrase en question et des étiquettes pré-terminales (des parties du discours).

Exemple 8 *(Parties du discours et catégories syntaxiques comme dans le corpus de TA-LaNa)*



Problème: Ambiguités d'attachement:

Paul regarde l'homme avec des jumelles

Le syntagme prépositionnel *avec des jumelles* peut faire partie du syntagme nominal de *homme* ou il peut être immédiatement dominé par le noeud S de toute la phrase.

Solutions possibles:

- sous-spécification: ne pas attacher le syntagme prépositionnel, indiquer peut-être les attachments possibles
- choisir un attachment en indiquant à l'utilisateur du corpus que cet attachment doit être interpreté comme sous-spécifié (par exemple choisir toujours l'attachment le plus haut)

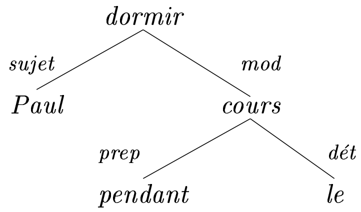
3.1.2 Structures de dépendances

Tesnière (1959) *Eléments de syntaxe structurale*

Mel'cuk (1988) *Dependency syntax: theory and practice*

Liens hiérarchiques entre mots. Ces liens expriment des relations syntaxiques de gouverneur et modificateur et ils ont des étiquettes de fonctions syntaxiques.

Exemple 9 *Une structure de dépendance syntaxique:*



Cela est une structure de dépendance pour

(a) Pendant le cours Paul dort
(b) Paul dort pendant le cours

Ce sont surtout la distinction complément/modificateur et les différents types de compléments qu'on veut annoter.

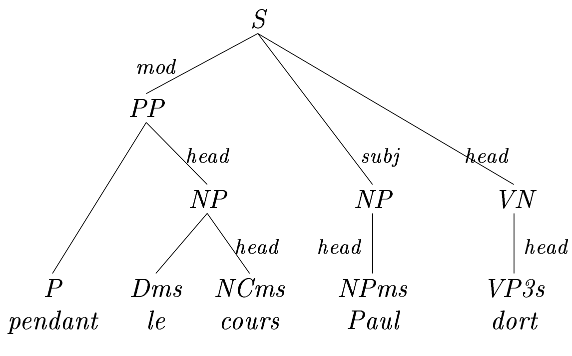
3.2 Comment annoter constituants et dépendances

Dans les corpus arborés (treebanks) on utilise en général des arbres pour annoter les structures de constituance.

Ces arbres sont enrichis avec des fonctions grammaticales.

Souvent, la structure des constituants et les fonctions grammaticales des constituants donnent déjà les dépendences qu'on aimerait annoter.

Exemple 10



Problème : les constituants discontinus

- dépendances non-bornées

  Que penses-tu que Jean va acheter aujourd'hui

  (le pronom interrogatif *que* est l'objet direct de *acheter*)

- Langues où l'ordre des mots est relativement libre comme l'allemand:

  das Fahhrad wollte er nicht reparieren
  la bicyclette voulait il ne pas réparer
  'il ne voulait pas réparer la bicyclette'

(*das Fahrrad* est l'objet direct de *reparieren*)

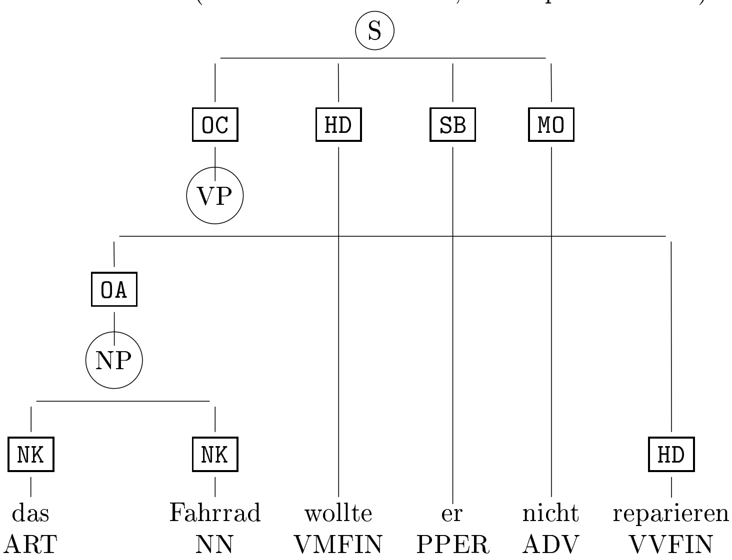
Evidemment, ces dépendances ne peuvent pas être annotées dans une structure d'arbre. Il faut une variante de structure de donnée ou un enrichissment des arbres de constituants.

Deux solutions:

- enrichir les arbres avec des constituants vides est des indices (comme dans le Penn Treebank)

```
(SBARQ (WHNP-1 Que)
       (SQ penses-tu
           (SBAR que
               (S (NP-SBJ Jean)
                  (VP va
                      (VP acheter
                          (NP *T*-1)
                          (ADVP aujourd'hui)))))))
```

- modifier les structures de donnée permises comme annotations en permettant des branches croisées (comme dans NEGRA, un corpus allemand).



Attention : le fait qu'un noeud est dessiné à gauche d'un autre noeud dans une structure à branches croisées ne veut pas nécessairement dire qu'il y a une précedence linéaire entre les deux. Le noeud VP par exemple n'est pas à gauche (das le sens de préc. lin.) du noeud décoré avec "wollte". Pour déterminer la position originale (dans le sens d'une annotation style Penn Treebank) d'un constituant comme le VP dans l'exemple NEGRA, on suppose en général que la tête (fonction `HD`=head) indique la position originale.

Un problème lié sont des verbes à contrôle et des verbes à montée ou une phrase a deux fonctions :

(a) *Jean veut venir*
(Jean sujet de veut et de venir)

(b) *Jean semble parler*
(Jean sujet de parler)

Dans un Penn Treebank style on a une catégorie vide à l'intérieur de la phrase avec l'infinitif. Dans NEGRA, on adopte des branches supplémentaires (secondary edges) pour ces deux exemples.