



# Cours sur le traitement automatique des langues (II)


Violaine Prince  
Université de Montpellier 2  
LIRMM-CNRS

# L'analyse syntaxique

- Le but : fournir une structure interprétable d'un texte.
  - ◆ si le langage est artificiel, il faut en plus vérifier la correction
- L'outil principal de l'analyse syntaxique
  - ◆ la grammaire
    - ✦ hors-contexte
    - ✦ dépendante du contexte
    - ✦ Sans restriction

# Rappel sur la hiérarchie de Chomsky

- Grammaires type 3 :
  - ◆ Grammaires régulières (langages de programmation)
- Grammaires de type 2 :
  - ◆ Grammaires hors contexte
    - ✦ Exemples d'application pour le langage naturel : avantages et inconvénients=> grammaires en FN Chomsky, grammaires logiques...

- 
- Grammaires de type 1:
    - ◆ Grammaires dépendantes du contexte
      - ✦ Grammaires transformationnelles
      - ✦ Certaines utilisations des LFG (lexical functional grammar), etc.
  - Grammaires de type 0
    - ◆ Grammaires totalement générales
      - ✦ Algorithme de réécriture de Markov
      - ✦ Puissance d'une machine de Turing



# Organisation de l'exposé


- Forme des règles de réécriture des différents types de grammaire dans le cadre du langage naturel
- Présentation des grammaires FN-Chomsky
- Equivalence entre des grammaires HC et des grammaires FN-Chomsky

# Pourquoi ?

- Les grammaires HC ont été historiquement très investies en TALN.
- La FN-Chomsky a le mérite d'être associée à un algorithme d'analyse en  $kn^3$  où  $k$  est la taille de la partie non terminale de la grammaire, et  $n$  le nombre de mots du fragment à analyser. ALGORITHME DE COCKE YOUNGER KASAMI

# Grammaires hors contexte

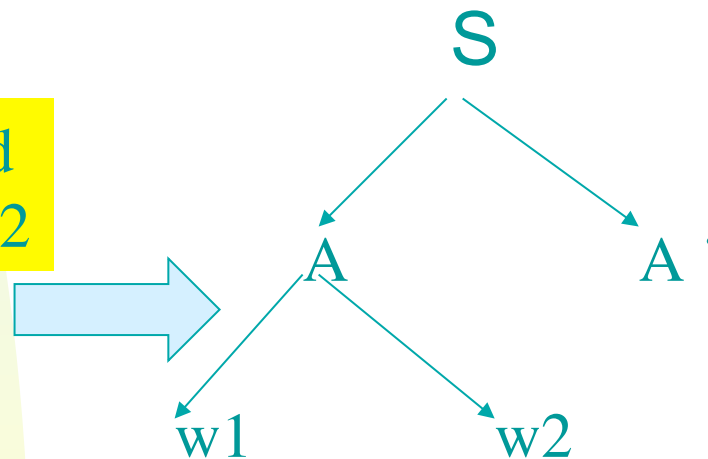
- Rappel sur les grammaires
  - ◆  $G = \{V_n, V_t, S, P\}$ 
    - ✦  $V_n$  : vocabulaire auxiliaire
    - ✦  $V_t$  : vocabulaire terminal
    - ✦  $S \in V_n$  : axiome
    - ✦  $P$  ensemble fini de productions
      - $p \in P$  si
      - $p \in (V_n \cup V_t) \times (V_n \cup V_t)$  et  $p = (w_1, w_2)$ .

- 
- Le système défini par  $\{Vt^*, S, P\}$  est un système formel
  - Le langage  $L$  engendré par  $G$  est l'ensemble des théorèmes du système formel
  - En langage naturel :
    - ◆  $V_n$  : ensemble des **catégories grammaticales** (étiquettes) et des catégories syntaxiques.
      - ✦ Ex : groupe verbal est une catégorie syntaxique. verbe est une catégorie grammaticale.



- ◆  $V_t$  : ensemble des mots de la langue (dictionnaire)
- ◆  $P$  : ensemble des règles de grammaire
- ◆  $S$  : axiome unitaire (phrase)
- ◆ Structure syntaxique : arborescence mémorisant la démonstration permettant d'obtenir un mot (dérivation).

correspond  
à  $A \rightarrow w_1 w_2$



# Hors contexte et dépendant du contexte

- G est dite hors contexte si, pour toute production du type
  - ◆  $A \rightarrow w, A \in V_n, \text{ et } w \in V_n \text{ ou } V_t.$
- G est dite dépendante du contexte si
  - ◆  $\alpha A \beta \rightarrow \alpha \gamma \beta \quad \alpha, \beta, \gamma \in V_n \cup V_t, A \in V_n$
- Exemples
  - ◆ une règle du type  $GN \rightarrow \text{Nom préposition Nom}$  est hors contexte
  - ◆ une règle du type « Le »  $GN \rightarrow \text{« Le » Adjectif Nom}$  est dépendante du contexte

# Grammaire non restreinte

- Toute règle de production de la forme :
  - ◆  $\alpha \rightarrow \beta$
- Équivalente à une machine de Turing

# Les grammaires normées de Chomsky

- Une grammaire normée de Chomsky est un ensemble de règles de réécriture (ou production) de la forme :
- $S \rightarrow S_1 S_2$ 
  - ◆ avec  $S \in V_n$ ,  $S_1 \in V_n$ ,  $S_2 \in V_n$
- Ou  $S \rightarrow A$ 
  - ◆ Avec  $S \in V_n$ ,  $A \in V_t$ .

# Les grammaires normées de Chomsky

- S, S1 et S2 sont des symboles non terminaux (dans l'ensemble  $V_n$  des étiquettes grammaticales et syntaxiques)
- A est un symbole terminal (dans l'ensemble  $V_t$  des lexèmes)
- Exemple :
  - ◆  $PH \rightarrow GN\ GV$
  - ◆  $GV \rightarrow V\ GN\ I\ V\ GNPREP$
  - ◆  $GN \rightarrow DET\ SN$
  - ◆  $GNPREP \rightarrow PREP\ GN$
  - ◆ Sont des règles valides sur des symboles non terminaux

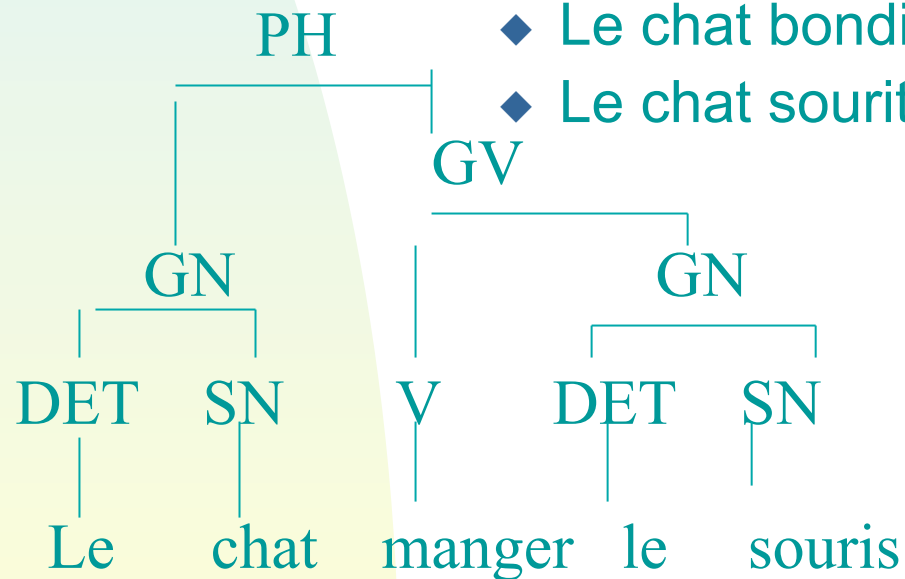
# Les grammaires normées de Chomsky

- Les règles suivantes permettent d'inclure des symboles terminaux :
- DET  $\rightarrow$  Le
- SN  $\rightarrow$  chat
- SN  $\rightarrow$  souris
- SN  $\rightarrow$  sourire
- V  $\rightarrow$  manger
- V  $\rightarrow$  bondir
- V  $\rightarrow$  sourire
- PREP  $\rightarrow$  sur
- PREP  $\rightarrow$  à

On supposera que l'analyse morphologique est capable de reconnaître les flexions et les conjugaisons.

# Les grammaires normées de Chomsky

- La grammaire ainsi définie, est capable de générer (ou de reconnaître) les phrases suivantes :
  - ◆ Le chat mange la souris.
  - ◆ Le chat bondit sur la souris.
  - ◆ Le chat sourit à la souris.



# Génération et analyse

- Les grammaires de Chomsky sont des grammaires génératives (puissance en génération)
- L'analyse est vue comme un cas particulier d'appariement entre :
  - ◆ Une phrase fournie
  - ◆ Une phrase que l'on peut engendrer par la grammaire
- En raison de l'ambiguïté des étiquettes, la combinatoire en analyse est élevée, et l'élimination des combinaisons « fausses » est une tâche importante.



# Combinatoire en analyse

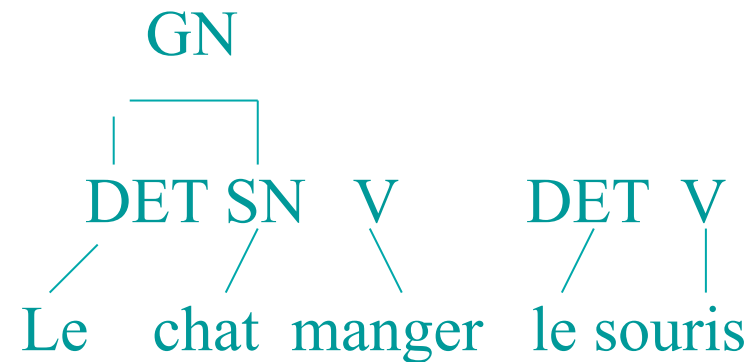
## ■ Exemple :

◆ Le chat mange la souris.



On cherchera à construire deux arbres

(arborescence précédente)



# Combinatoire en analyse

- On dit que l'analyse a réussi si et seulement si, à partir d'une phrase donnée, on construit par les règles de grammaire une arborescence de nœud PH (racine).
- Dans l'exemple précédent, avec la valeur V pour « souris » la phrase ne mène pas à une arborescence de nœud PH. Elle doit donc être éliminée.

# Difficultés et « backtracking »

- Plus la grammaire est importante, plus la combinatoire est élevée.
- Exemple :
  - ◆ Supposons que l'on rajoute la règle terminale :
    - ✦  $GN \rightarrow Le$
  - ◆ Pour rendre compte des pronoms personnels compléments d'objet.

# Exemple de construction d'arborescences

Le chat mange la souris.

DET SN V DET SN  
GN GN V

6 arborescences seront construites avec comme base

DET SN V DET SN (c'est la bonne)

GN SN V DET SN  $\mid$  = GN SN V GN  $\mid$  = GN SN GV

DET SN V GN SN  $\mid$  = GN V GN SN  $\mid$  = GN GV SN  $\mid$  = PH SN

DET SN V GN V  $\mid$  = GN V GN V  $\mid$  = GN GV V  $\mid$  = PH V

GN SN V GN SN  $\mid$  = GN SN GV SN

GN SN V GN V  $\mid$  = GN SN GV V

GN SN V DET V

# Algorithme d'analyse

- Il existe un algorithme, nommé algorithme de Cocke-Younger-Kasami qui permet de savoir si une phrase est analysable par une grammaire normée de Chomsky, avec une complexité en  $O(n^3)$ . Fourni en TD.
- Une phrase peut ne pas être analysable par une grammaire de Chomsky parce que:
  - ◆ La grammaire est non couvrante
  - ◆ La phrase n'est pas grammaticale (mal formée).
  - ◆ Il y a des mots inconnus.

# Théorème d'équivalence

- Toute grammaire de type réécriture, hors contexte, peut être réécrite sous forme de grammaire normée de Chomsky.
- Exemple :
  - ◆  $S \rightarrow S1 S2 S3$
  - ◆  $S3 \rightarrow S5$
  - ◆  $S2 \rightarrow k$
- Peut se réécrire sous forme :
  - ◆  $S \rightarrow S1 S4$
  - ◆  $S4 \rightarrow S2 S3$
  - ◆  $S2 \rightarrow k$

# Intégrité de la grammaire

- Une grammaire est intègre en génération si tout symbole non terminal peut se réduire, par une série de réécritures, en un (ou plusieurs) symbole(s) terminal(ux).
- Pas de symboles « vides ».
  - ◆ Ex:  $GV \rightarrow V \text{ GNPREP}$
  - ◆  $\text{GNPREP} \rightarrow \text{PREP GN}$
  - ◆ Il faut qu'il existe une règle terminale avec  $\text{PREP} \rightarrow \text{xxx}$



# Les grammaires normées de Chomsky

- Pour le langage naturel :
  - ◆ Analyse en constituants
- Des défauts :
  - ◆ Rechercher une grammaire couvrante, intègre.
  - ◆ Ne traite pas les phrases agrammaticales
  - ◆ Multiplication des règles de génération pour les phrases particulières (nominales, subordonnées, relatives, etc...).



# Constituants et dépendances

- Un constituant est un élément de construction syntaxique.
- La catégorie grammaticale est un constituant atomique :
  - ◆ Ex : nom, adjectif, verbe, pronom, adverbe, locution adverbiale, déterminant...
- Les constituants non atomiques sont :
  - ◆ Groupe nominal, adjectival, verbal, prépositionnel
  - ◆ Phrase nominale
  - ◆ Proposition principale, relative, etc...

# Analyse en constituants

- Grammaire de constituants en FN-Chomsky :
  - ◆  $PH \rightarrow GN, GV$
  - ◆  $GN \rightarrow DET, GN$
  - ◆  $GN \rightarrow ADJ, NOM$
  - ◆  $GN \rightarrow NOM, ADJ$
  - ◆  $GV \rightarrow V GN$

Peut générer des phrases de la forme :

« Le petit chat mange la souris grise »

# Terminologie

- La notion de P.O.S (part-of-speech) tagging est intermédiaire entre la détermination de la catégorie grammaticale (constituant atomique) et un constituant de faible granularité (par exemple un GN de la forme DET NOM)
- La notion de « chunk » correspond de manière un peu floue à des constituants de plus grande granularité et mélange les deux notions de constituants et de dépendance.

# Analyse en dépendance

- Une dépendance définit l'importance d'un constituant dans la composition de la phrase.
- Son rôle est intermédiaire entre la syntaxe et la sémantique.
- Théorie de l'effacement :
  - ◆ Certains éléments sont « obligatoires » pour que la phrase soit construite correctement :
    - ♦ sujet,
    - ♦ Prédicat.

- D'autres éléments sont facultatifs :
  - ◆ Les compléments d'objet
  - ◆ Les compléments circonstanciels
- Récurrence de l'aspect « obligatoire » vs « facultatif » à tous les niveaux des constituants :
  - ◆ Dans un groupe nominal, le nom, ou le pronom peuvent être indispensables
  - ◆ Le déterminant n'est indispensable que dans le cas d'un nom commun (en Français)
  - ◆ L'adjectif est facultatif
  - ◆ Dans une composition complexe de la forme :
    - ✦ Nom Préposition Nom
    - ✦ Le premier nom ne peut pas être enlevé (sauf dans certains cas).

# La notion de « gouvernement »

- Théorie du gouvernement de Chomsky
- Théorie de la dépendance chez Tesnière
- Notion de tête dans les grammaires de dépendance (e.g., grammaires HPSG)
- Idée de base : il existe un constituant atomique ou non, qui gouverne à l'intérieur de la structure à laquelle il contribue.

# Exemples de gouvernement

- Le petit chat
- Le médecin de famille
- Manger proprement.
- Dormir dans son lit.
- Le petit chat dort dans son panier.
- Hier, le médecin de famille est venu chez nous.
- [Le médecin que ma sœur m'a recommandé] a sauvé [la vie de mon père].

# Qu 'est-ce qui gouverne une phrase ?

- Son prédicat verbal : approche logique
  - ◆ Dort (Chat, Panier)
  - ◆ Venir (Médecin de famille, chez nous, hier)
- Son prédicat et le sujet de ce dernier : approche linguistique

Ph -> Sujet Groupe verbal

Est une phrase bien formée.

Mais le prédicat peut ne pas être verbal, et le sujet, pas nominal :

Dormir, quel délice !



# Une grammaire de constituants et de dépendances ?

- Ph -> Sujet Groupe Verbal
- Groupe Verbal -> GV COD
- Groupe Verbal -> GV Complément Circ
- Groupe Verbal -> GV COI
- COD -> GN
- Complément Circ -> PREP Pronom
- COI -> GN PREP GN
- GN -> DET GN
- GN -> ADJ NOM

- Le mélange est possible mais :
  - ◆ Difficile d'écrire proprement une FN-Chomsky
  - ◆ Beaucoup de dépendances sont sensibles au contexte (on sort du cadre du « context-free »)
    - ✦ Les compléments rejetés en début de phrase
    - ✦ Le non effacement des compléments :
      - Je suis allé à Paris
      - Je suis allé
  - ◆ La détermination du gouvernement est souvent ambiguë.
  - ◆ La détermination de la nature du complément nécessite une connaissance sémantique

# Les grammaires normées de Chomsky

- Pour une première approche des problèmes peut servir d'accroche « pédagogique » :
  - ◆ Implémentation de l'algorithme de Cocke
  - ◆ Recherche de la couverture grammaticale d'un ensemble de phrases données (bien formées).
- Marche mal sur du tout venant, et sur des grandes masses de données.