Grammaire, grammaire régulière

Intelligence Artificielle et Systèmes Formels
Master 1 I2L

Sébastien Verel
verel@lisic.univ-littoral.fr
http://www-lisic.univ-littoral.fr/~verel

Université du Littoral Côte d'Opale
Laboratoire LISIC
Equipe CAMOME

Objectifs de la séance 05

- Connaître la définition d'une grammaire
- Savoir dessiner l'arbre de dérivation d'un mot selon une grammaire
- Connaitre la définition d'une grammaire régulière
- Savoir concevoir un automate fini reconnaissant le langage d'une grammaire régulière
- Savoir définir une grammaire engendrant le même langage qu'un automate fini déterministe
- Connaitre la classification de Chomsky des langages

Plan

Là où nous en sommes

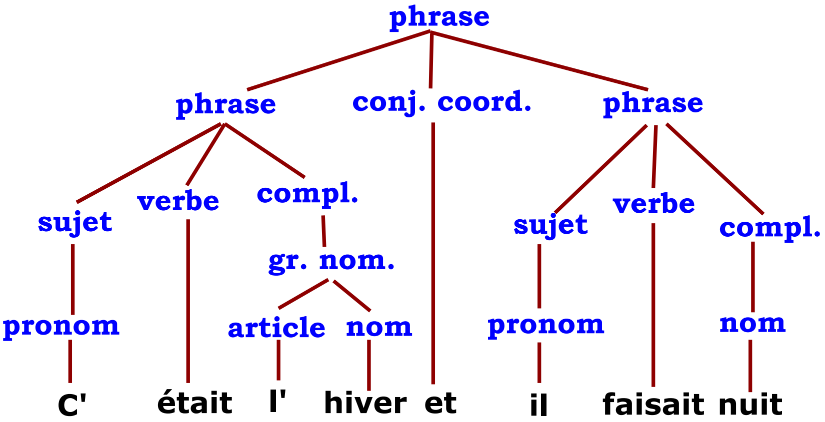
- Les mots d'un langage rationnel sont reconnus par des automates finis et réciproquement.
- Les automates sont des machines abstraites capables de réaliser des calculs sur des mots.
- Les notions de langage et de calcul sur machine sont très proches :
  - Pour tout langage rationnel, il existe un automate reconnaissant ce langage
  - Tout automate reconnait un langage rationnel

Vers où l'on va

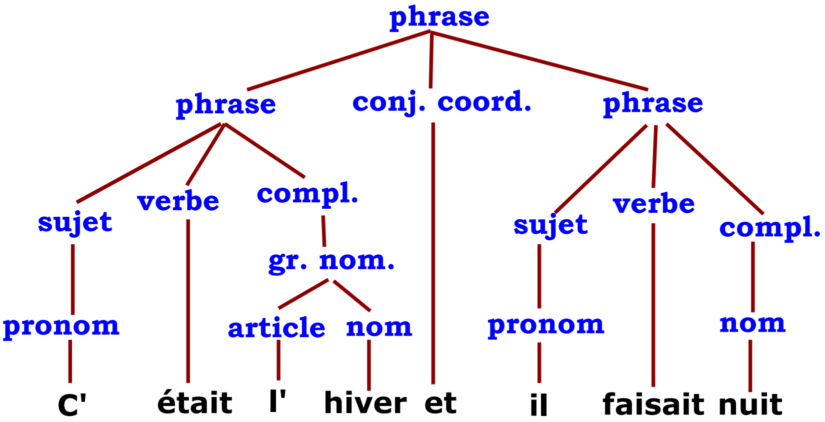
- Il est possible de définir d'autres propriétés et d'autres types de machines abstraites qui permettent de définir d'autres classes de langages.
  L'expressivité du langage et la capacité de calcul de la machine sont alors différentes.
- Les questions que l'on se pose sont alors les mêmes :
  - mode de lecture,
  - description formelle/algébrique langage,
  - équivalence avec d'autres classes de langages,
  - capacité de la machine reconnaissant le langage.

Par exemple, on peut remplacer expression régulière par grammaire ou automate par machine de Turing...

Exemple littéraire



Exemple littéraire



Première phrase de "La position du tireur couché", J.-P. Manchette.

Intuition : la réécriture

Réécriture (source CNRTL)

Règle permettant de retranscrire, suivant un principe de transformation, une suite de symboles en une autre.

Intuition : la réécriture

Réécriture (source CNRTL)

Règle permettant de retranscrire, suivant un principe de transformation, une suite de symboles en une autre.

Moyen expressif puissant

Au lieu de définir un langage
par les mots vérifiant une expression algébrique,
on utilise un ensemble de fonctions de l'ensemble des mots dans lui-même, et plus généralement des couples de mots.

Comment peut-on réécrire ?

Intuition

- Objet initial : S
- Règles de réécriture, de production :

  "alpha se réécrit en beta"
  $$\alpha \longrightarrow \beta$$

- Relation de dérivation : $\vdash$
  - si $\alpha \longrightarrow \beta$ alors $\alpha \vdash \beta$
  - si $\alpha \longrightarrow \beta$ et
    si en remplaçant α par β dans γ on obtient δ
    alors $\gamma \vdash \delta$
- Langage engendré : $\{\varphi \ : \ S \vdash^* \varphi\}$

Exemple graphique

Alphabet de symboles : { F, -, + }
Règles de réécriture (ou de production) :
F → F-F+F+FF-F-F+F (P1)

Cette règle peut s'interpréter graphiquement. Sur un quadrillage,

- F signifie "avancer" d'une unité,
- - "tourner" à droite de 90 degrés,
- + "tourner" à gauche de 90 degrés.

Questions

Graphiquement,

- Dériver F en appliquant 1 fois (P1)
- Dériver F-F-F-F en appliquant plusieurs (P1)

Réécriture avec une production

Soit Σ un alphabet.

Production

Une **production** est un couple (α, β) de $\Sigma^+ \times \Sigma^*$, noté :

$$\alpha \rightarrow \beta$$

α est appelé prédécesseur et β successeur.

Règle de réécriture

La règle de réécriture $\vdash_{\alpha\rightarrow\beta}$ est une relation binaire définie par :
Pour tous les mots v et w sur Σ,

$$v \vdash_{\alpha\rightarrow\beta} w$$

ssi

il existe 2 mots v_1 et v_2 sur Σ

tels que $v = v_1\alpha v_2$ et $w = v_1\beta v_2$.

Grammaire

Définition

Une grammaire est un quadruplet (N, Σ, P, S) où :

- N est un ensemble fini de symboles **non terminaux**
- Σ est un ensemble de symboles **terminaux**
- P est un ensemble fini de **productions**

$$\{\varphi \to \psi \;:\; \varphi \in (N \cup \Sigma)^+ \text{ et } \psi \in (N \cup \Sigma)^*\}$$

- $S \in N^+$ est l'**axiome** (mot initial)

Propriétés

- Vocabulaire : $V = N \cup \Sigma$
- $N \cap \Sigma = \emptyset$

Notation pratique

Notation d'un ensemble de productions

Plusieurs productions avec la même partie gauche peuvent s'écrire en une seule avec le symbole "ou" $|$:

$$\begin{cases} \alpha \rightarrow \beta_1 \\ \alpha \rightarrow \beta_2 \end{cases}$$

peut se noter

$$\alpha \rightarrow \beta_1 \mid \beta_2$$

Dérivation

Dérivation

$(w_0, w_1, \ldots, w_n)$ est une dérivation de v en w selon la grammaire $G = (N, \Sigma, P, S)$ ssi $w_0 = v$ et $w_n = w$ et pour tout $i \in \{0, n-1\}$ il existe une production $(\alpha, \beta) \in P$ telle que $w_i \vdash_{\alpha \to \beta} w_{i+1}$.

Remarque : les productions s'appliquent successivement.

Relation $\vdash^*$

$v \vdash^* w$ s'il existe une dérivation entre les mots v et w.

Remarque : La relation $\vdash^*$ est une relation binaire transitive.

Calcul, langage engendré

Langage engendré

Le langage engendré par la grammaire $G = (N, \Sigma, P, S)$ est :

$$L(G) = \{\varphi \in \Sigma^* \ : \ S \vdash^* \varphi\}$$

Exemple (1)

Soit la grammaire $G_1 = (N, T, P, A)$ avec :

- $N = \{A, B, C\}$
- $T = \{a, b\}$
- $$P = \{ \begin{array}{lcl} A & \rightarrow & aB, \\ B & \rightarrow & bC, \\ C & \rightarrow & aC \\ C & \rightarrow & bC \\ C & \rightarrow & \epsilon\} \end{array}$$

Quel est le langage engendré par G_1 ?

Exemple (2)

Soit la grammaire $G_2 = (N, T, P, A)$ avec :

- $N = \{A, B, O, T\}$
- $T = \{o, t, p\}$
- $$P = \{ \begin{array}{lcl} A & \rightarrow & tA, \\ A & \rightarrow & oA, \\ A & \rightarrow & pA, \\ A & \rightarrow & tT, \\ T & \rightarrow & oO, \\ O & \rightarrow & pB, \\ B & \rightarrow & \epsilon, \\ B & \rightarrow & tB, \\ B & \rightarrow & oB, \\ B & \rightarrow & pB\} \end{array}$$

Quel est le langage engendré par G_2 ?

Exemple (3)

Soit la grammaire $G_3 = (N, T, P, A)$ avec :

- $N = \{A\}$
- $T = \{0, 1\}$
- $$P = \{ \begin{array}{lll} A & \rightarrow & 0A1, \\ A & \rightarrow & \epsilon\} \end{array}$$

Quel est le langage engendré par G_3 ?

Exemple (3)

Soit la grammaire $G_3 = (N, T, P, A)$ avec :

- $N = \{A\}$
- $T = \{0, 1\}$
- $$P = \{ \begin{array}{lcl} A & \rightarrow & 0A1, \\ A & \rightarrow & \epsilon\} \end{array}$$

Quel est le langage engendré par G_3 ?

Attention tous les langages ne sont pas des langages rationnels !
Et loin de là, pensez au cardinal de ces ensembles...

Arbre de dérivation

Important : Soit $G = (N, T, P, A)$ grammaire dont les productions n'ont qu'un seul symbole non-terminal comme prédécesseur.

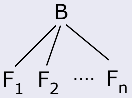
Arbre de dérivation

Représentation graphique de la dérivation d'un mot depuis l'axiome.

Arbre de dérivation du mot w engendré par G

Arbre tel que :

- Racine étiquetée par le symbole initial S
- Feuilles étiquetées par les éléments de $T \cup \{\epsilon\}$
- Noeuds internes étiquetés par les éléments de N



  si une production $B \rightarrow F_1 F_2 \ldots F_n$

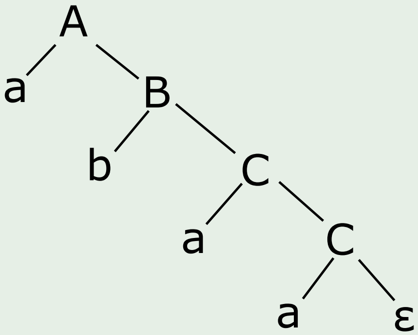
- w est formé par la concaténation des feuilles obtenues par un parcours en profondeur de l'arbre

Exemple

Arbre de dérivation de $w = abaa$ engendré par la grammaire G_1

Exemple

Arbre de dérivation de $w = abaa$ engendré par la grammaire G_1

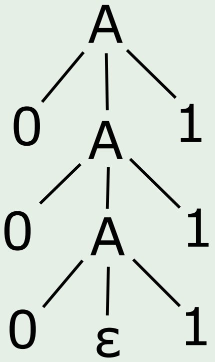


Exemple

Arbre de dérivation de $w = 000111$ engendré par la grammaire G_3

Exemple

Arbre de dérivation de $w = 000111$ engendré par la grammaire G_3



Définition

Grammaire régulière à droite

Une grammaire $G = (N, T, P, A)$ est régulière à droite si toutes les productions sont de la forme :

$$\mathrm{B} \rightarrow \mathrm{aC} \quad \text{ou} \quad \mathrm{B} \rightarrow \mathrm{a} \quad \text{ou} \quad \mathrm{B} \rightarrow \epsilon$$

avec $\mathrm{B} \in \mathrm{N}$, $\mathrm{C} \in \mathrm{N}$ et $\mathrm{a} \in \mathrm{T}$.

Définition

Grammaire régulière à droite

Une grammaire $G = (N, T, P, A)$ est régulière à droite si toutes les productions sont de la forme :

$$\mathrm{B} \rightarrow \mathrm{aC} \quad \text{ou} \quad \mathrm{B} \rightarrow \mathrm{a} \quad \text{ou} \quad \mathrm{B} \rightarrow \epsilon$$

avec $\mathrm{B} \in \mathrm{N}$, $\mathrm{C} \in \mathrm{N}$ et $\mathrm{a} \in \mathrm{T}$.

Grammaire régulière à gauche

Une grammaire $G = (N, T, P, A)$ est régulière à gauche si toutes les productions sont de la forme :

$$\mathrm{B} \rightarrow \mathrm{Ca} \quad \text{ou} \quad \mathrm{B} \rightarrow \mathrm{a} \quad \text{ou} \quad \mathrm{B} \rightarrow \epsilon$$

avec $\mathrm{B} \in \mathrm{N}$, $\mathrm{C} \in \mathrm{N}$ et $\mathrm{a} \in \mathrm{T}$.

Exemple

Question

Parmi les grammaires G_1, G_2 et G_3,
lesquels sont des grammaires régulières ?

Lien AFD / grammaire régulière

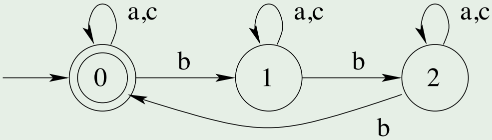
Théorème

Un langage L est rationnel
si et seulement si
il existe une grammaire régulière qui engendre les mots de L.

cf. exercice 1 Fiche 05

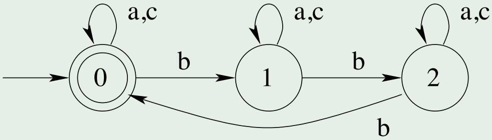
Construction d'une grammaire régulière à partir d'un AFD

Exemple



Construction d'une grammaire régulière à partir d'un AFD

Exemple



- $T = \{a, b, c\}$
- $P = \{ \quad A \rightarrow aA, \quad A \rightarrow cA, \quad A \rightarrow bB,$
  $B \rightarrow aB, \quad B \rightarrow cB, \quad B \rightarrow bC,$
  $C \rightarrow aC, \quad C \rightarrow cC, \quad C \rightarrow bA,$
  $A \rightarrow \epsilon\}$
- $N = \{A, B, C\}$

Construction d'une grammaire régulière à partir d'un AFD

Soit un automate $\mathcal{A} = (\Sigma, Q, T, q_0, F)$
La grammaire $G = (N, T, P, A)$ telle que :

- $N = Q$
- $T = \Sigma$
- $A = q_0$
- dont les productions P sont :
  $q_i \rightarrow aq_j$ lorsque l'automate a pour transition $T(q_i, a) = q_j$
  et $q_i \rightarrow \epsilon$ pour tout $q_i \in F$

engendre le même langage que l'automate $\mathcal{A}$.

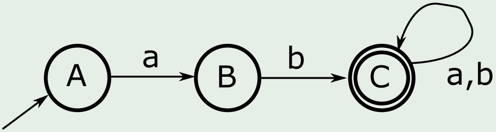
Remarque : il y a des renommages implicites dans la construction

Construction d'un AFD à partir d'une grammaire régulière

- $N = \{A, B, C\}$
- $T = \{a, b\}$
- $$P = \{ \begin{array}{lcl} A & \rightarrow & aB, \\ B & \rightarrow & bC, \\ C & \rightarrow & aC \\ C & \rightarrow & bC \\ C & \rightarrow & \epsilon\} \end{array}$$

Construction d'un AFD à partir d'une grammaire régulière

- $N = \{A, B, C\}$
- $T = \{a, b\}$
- $$P = \{ \begin{array}{lcl} A & \rightarrow & aB, \\ B & \rightarrow & bC, \\ C & \rightarrow & aC \\ C & \rightarrow & bC \\ C & \rightarrow & \epsilon\} \end{array}$$



Construction d'un AFD à partir d'une grammaire régulière

Soit la grammaire régulière à droite $G = (N, T, P, A)$.
L'automate $\mathcal{A} = (\Sigma, Q, T, q_0, F)$ telle que :

- $Q = N \cup \{X\}$
- $\Sigma = T$
- $q_0 = A$
- dont les transitions sont :
  $T(q_i, a) = q_j$ pour les productions $q_i \rightarrow aq_j$
  $T(q_i, a) = X$ pour les productions $q_i \rightarrow a$
- $F = \{X\} \cup \{B \ : \ \text{production } B \rightarrow \epsilon\}$

reconnait le même langage que la grammaire G.

Remarque : il y a des renommages implicites dans la construction

Langages algébriques ou non-contextuels

Grammaire algébrique

Une grammaire $G = (N, \Sigma, P, S)$ est **non-contextuelle** (context-free in english), ou algébrique, si les productions sont de la forme :

$$B \rightarrow w \text{ avec } B \in N \text{ et } w \in (N \cup \Sigma)^*$$

Langage algébrique

Un langage L est algébrique
s'il existe une grammaire algébrique G telle que $L(G) = L$.

Remarque : Tout langage rationnel est algébrique, mais la réciproque est fausse.
Exemple : les langages de programmation classiques

Hierarchie de Chomsky

Classification de Chomsky

Type	Langages	Grammaires
3	Rationnels ou réguliers	régulières à droite $A \longrightarrow a,\ A \longrightarrow aB,\ A \longrightarrow \epsilon$ $A, B \in N \ a \in T$ (régulières à gauche)
2	algébriques ou non-contextuels	algébriques, non-contextuelles $A \longrightarrow \alpha$ $A \in N \ \alpha \in (N \cup T)^*$
1	contextuels	contextuelles, monotones $\alpha \longrightarrow \beta$ ou $A \longrightarrow \epsilon$ $\alpha, \beta \in (N \cup T)^*$, A axiome $\lvert\alpha\rvert \leq \lvert\beta\rvert$
0	récursivement énumérables	contextuelles avec effacement $\alpha \longrightarrow \beta$ $\alpha \in (N \cup T)^+ \ \beta \in (N \cup T)^*$ aucune contrainte

Formes normales

- Pour comparer les grammaires, on peut les écrire sous forme normalisée
- Il existe plusieurs formes normales (Chomsky, Greibach, etc.)
- Il existe des algorithmes pour transformer les grammaires en forme normale

Forme normale de Chomsky

Une grammaire $G = (N, \Sigma, P, S)$ est sous forme normale de Chomsky si toutes les productions sont de la forme :

- $A \rightarrow a$
- $A \rightarrow BC$
- $S \rightarrow \epsilon$

avec $A, B, C \in N$ et $a \in \Sigma$ (et si on a $S \rightarrow \epsilon$, S ne figure dans aucun membre droit d'une autre production)

Conclusion

Langages rationnels ou réguliers

- Langages clos par : union, produit, étoile,
- Ensemble dénombrable de langage,
- Caractérisés par les expressions régulières
- Reconnus par les automates finis (déterministe ou non)
- Engendrés par les grammaires régulières (à droite ou gauche)