

Introduction à la compilation

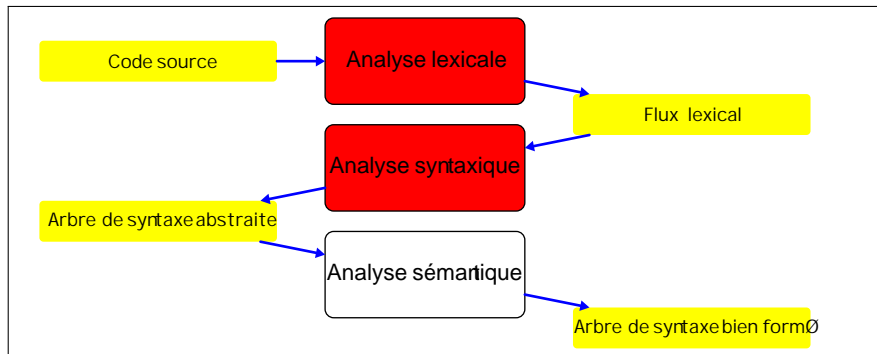
Cours 2 : Analyse lexicale et syntaxique

Yann Røgis-Gianas

y

PPS - Université Denis Diderot Paris 7

Rappel du dernier cours



Première partie du cours :

L'analyse lexicale et l'analyse syntaxique dans la partie avant d'un compilateur

Vue d'ensemble

Grandes lignes

- | L'analyse lexicale traduit un **flux de caractères** en un **flux de lexèmes**.
- | L'analyse syntaxique traduit un **flux de lexèmes** en un **arbre de syntaxe abstraite**.
-) Dans les deux cas, il s'agit **d'exhiber la structure implicite d'un flux**.
 - | Cette structure nous servira à donner un **sens** aux programmes.
-) Pour le moment, nous nous focalisons sur les problèmes suivants :
 1. Comment spécifier la structure de tous les programmes d'un langage ?
 2. Comment reconnaître la structure d'un programme et la reconstruire ?

Décrire une syntaxe

| Pour voir un programme comme un arbre, on doit répondre à ces questions :

- | Quelles sont ses feuilles ? Les symboles **terminaux**.
-) les mots-clés du langage, sa ponctuation, ses atomes de sens.
- | Quels sont ses nœuds ? Les symboles **non terminaux**.
-) ses **catégories syntaxiques** (expressions, déclarations, types, ...)

Décrire une syntaxe : exemple du langage

Terminaux :

- ┆ Mots-clés : sum, from, to, do, def.
- ┆ Identifiants : x, y, z, foo, bar, ...
- ┆ Constantes entières : 0, 1, 2, 3, ...
- ┆ Symboles : '+', '*', '/', '-', '(', ')', ',', '=', ...

Non terminaux :

- ┆ Expression : les expressions arithmétiques formées :
 - ┆ par application des symboles d'opérateurs appliqués à des sous-expressions respectant les priorités entre ces symboles ;
 - ┆ par application de l'opérateur sum suivi d'un identifiant, du mot-clé from, d'une expression, du mot-clé to, d'une expression, du mot-clé and et d'une expression ;
 - ┆ par application d'une fonction à une liste d'expressions, séparées par des virgules et entourée de parenthèses.
- ┆ Déclaration de fonctions : le mot-clé def suivi d'un identifiant, d'une liste d'identifiants entre parenthèses et séparés par des virgules, du symbole '=' et d'une expression.
- ┆ Programme : une liste de déclarations de fonction suivie d'une expression.

) Une description peu concise ...et incomplète!

(Quelles sont les priorités respectives des symboles? Comment les décrire?)

Grammaires

Une **grammaire** (g n rative, syntagmatique, formelle, ...) est un quadruplet $(N; T; R; S)$ o  :

- | N et T sont des ensembles de symboles disjoints.
- | R est un ensemble de paires (P, Q) avec $P \in (N \cup T)^+$ et $Q \in (N \cup T)^*$
- | Il existe un symbole $S \in N$.

Les r gles de la grammaire (P, Q) de R , sont en g n ral not es $P \rightarrow Q$.

- | N est l'ensemble des symboles non-terminaux.
- | T est l'ensemble des symboles terminaux.
- | R est l'ensemble des r gles.
- | S est le symbole d'entr e.

1. A^+ repr sente l'ensemble des s quences non vides d' l ments de A .
2. A^* repr sente l'ensemble des s quences vides d' l ments de A .

La grammaire du langage Martha

- | Keywords fsum; from; to; do; defg
- | f; x 2 Identifiers fa;:::;zg⁺ n Keywords²
- | i 2 Integers f0;:::;9g⁺
- | T f'+';'*' '/' '/' '-' '-' '(' ')' ':' ';' '=' g [Keywords [Identifiers
- | N
- | fExpression; Factor; Term; SExpressionsg [
- | fSIdentifiers; De nition; De nitions; Programg
- | R est :

Expression	!	Expression '+' Term
Expression	!	Expression '-' Term
Expression	!	Term
Term	!	Term '*' Factor
Term	!	Term '/' Factor
Term	!	Factor
Factor	!	i
Factor	!	x
Factor	!	f '(' SExpressions ')'
Factor	!	'(' Expression ')'

La grammaire du langage Martha (suite)

SExpressions	!	Expression
SExpressions	!	Expression ' , ' SExpressions
SIdenti ers	!	x
SIdenti ers	!	x ' , ' SIdenti ers
De nition	!	def f '(' SIdenti ers ') ' '=' Expression
De nitions	!	
De nitions	!	De nition De nitions
Program	!	De nitions Expression

La grammaire du langage Martha , écriture condensée

Expression	!	Expression '+' Term j Expression '-' Term j Term
Term	!	Term '*' Factor j Term '/' Factor j Factor
Factor	!	i j x j f '(' SExpressions ')' j '(' Expression ')'
SExpressions	!	Expression j Expression ',' SExpressions
SIdentifiers	!	x j x ',' SIdentifiers
Definition	!	def '(' SIdentifiers ')' '=' Expression
Definitions	!	j Definition Definitions
Program	!	Definitions Expression

- | On peut déduire T et N par simple observation des règles.
- | On utilise le symbole '|' pour noter une alternative dans la grammaire.

La grammaire du langage ~~Ma~~ Martha

Exercice

-) Modifiez la grammaire pour accepter la définition et l'utilisation de fonctions d'arité nulle.³

3. C'est-à-dire dont le nombre d'arguments formels vaut zéro.

Grammaires génératives en OCaml

- On peut donner un type OCaml à ces objets :

```
type ('t, 'n) symbol =  
  j Terminal of 't  
  j NonTerminal of 'n
```

```
type ('t, 'n) lhs =  
  ('t, 'n) symbol list
```

```
type ('t, 'n) rhs =  
  ('t, 'n) symbol list
```

```
type ('t, 'n) rule =  
  ('t, 'n) lhs ('t, 'n) rhs
```

```
type ('t, 'n) grammar =  
  ('t, 'n) rule list
```

Grammaire de Martha en OCaml

```
type ('t, 'n) rule_alternative = 'n ! ('t, 'n) rule list

(* ('t, 'n) rhs ! ('t, 'n) rule_alternative *)
let (!) rhs =
  fun lhs ! [((NonTerminal lhs, []), rhs)]

(* ('t, 'n) rule_alternative ! ('t, 'n) rhs *)
(* ! ('t, 'n) rule_alternative *)
let ( j ) rhs1 rhs2 =
  fun lhs ! (rhs1 lhs @ (! rhs2) lhs)

(* ('t, 'n) rule_alternative ! 'n ! ('t, 'n) rule list *)
let ( ! ) lhs rhs = rhs lhs
```

Quelques notations.

Terminaux et non-terminaux en OCaml

```
type terminal =  
  j Int  
  j Identifier  
  j Lparen j Rparen j Comma  
  j Plus j Minus j Star j Slash  
  j Sum j From j To j Do j Def
```

```
type nonterminal =  
  j Expression  
  j Term  
  j Factor  
  j SExpressions  
  j SIdentifiers  
  j Definition  
  j Definitions  
  j Program
```

Grammaire de Martha en OCaml

```
let martha = [  
  Expression !  
  ! [ NonTerminal Expression ; Terminal Plus ; NonTerminal Term ]
```

Produire un mot à partir d'une grammaire

- En suivant les règles de la grammaire de Martha, on peut engendrer un programme syntaxiquement correct de ce langage :

Forme intermédiaire	Règle utilisée
Program	
Definitions Expression	Program ! Definitions Expression
Expression	Definitions !
Expression '+' Term	Expression ! Expression '+' Term
Expression '+' Factor	Term ! Factor
Expression '+' 21	Factor ! i
Term '+' 21	Expression ! Term
Factor '+' 21	Term ! Factor
21 '+' 21	Factor ! i

Produire un mot à partir d'une grammaire

- On aurait pu obtenir le même programme par un chemin différent :

Forme intermédiaire	Règle utilisée
Program	
Définitions Expression	Program ! Définitions Expression
Expression	Définitions !
Expression '+' Term	Expression ! Expression '+' Term
Term '+' Term	Expression ! Term
Term '+' Factor	Term ! Factor (à l'occurrence 2)
Term '+' 21	Factor ! i
Factor '+' 21	Term ! Factor
21 '+' 21	Factor ! i

(Notons que l'occurrence précise d'application d'une règle peut être nœc

Les règles en toute généralité

- | Rien ne nous empêche d'écrire une règle de la forme :
Identifier ', ' Identifier ') ' ! Identifier and Identifier ') '
-) En utilisant plusieurs symboles à la gauche d'une règle, on peut définir un **contexte** d'application de cette règle.

Dérivation

- | Soit une grammaire $G = (T; N; R; S)$
- | Soient u et v , deux séquences de T^* .
- | On écrit $u \Rightarrow v$, qui se lit u dérive v , si :
 - | la séquence u s'écrit $w_1 i_1 \dots i_m w_2$;
 - | la séquence v s'écrit $w_1 p_1 \dots p_n w_2$;
 - | la règle $i_1 \dots i_m \rightarrow p_1 \dots p_n$ est dans R .
- | En d'autres termes, u est dont une sous-séquence a été réécrite par une règle de R .
- | La relation \Rightarrow est appelée **relation de dérivation immédiate** de G .
- | La fermeture transitive et réflexive de cette relation est appelée **relation de dérivation** de G . Elle est notée \Rightarrow^* .
- | Le langage engendré par G que nous noterons généralement est :

$$\{v \in T^* \mid S \Rightarrow^* v\}$$

Dérivation en Caml

```
(* Une application de règle est déterminée par *)  
(* son lieu d'application et la règle appliquée. *)  
type ('t, 'n) rule_application =  
  int * ('t, 'n) rule  
  
(* Une dérivation peut être représentée par *)  
(* une liste d'applications de règle. *)  
type ('t, 'n) derivation =  
  ('t, 'n) rule_application list
```

Dérivation en Caml

```
let rulenumber g index =  
  List.nth g index  
  
let rulesabout (g : ('t, 'n) grammar) t =  
  List.iter function (  
    j (NonTerminal rt, []), _) rt = t  
    j _ ! false)  
    g  
  
let ruleabout g?(index=0) t =  
  List.nth (rulesabout g t) index
```

Quelques accesseurs de règle dans une grammaire.

Dérivation en Caml

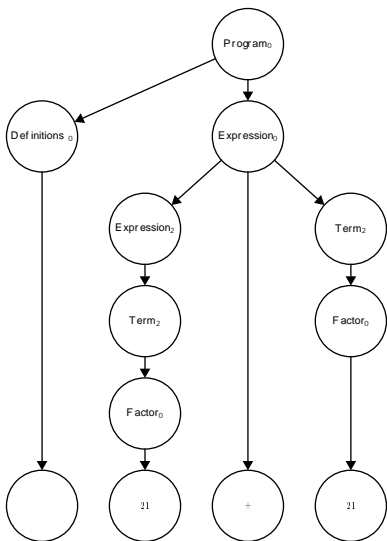
```
let derivationexample = [  
  0, ruleabout martha Program;  
  0, ruleabout martha index 0 Définitions;  
  0, ruleabout martha index 0 Expression;  
  2, ruleabout martha index 2 Term;  
  2, ruleabout martha index 0 Factor;  
  0, ruleabout martha index 2 Expression;  
  0, ruleabout martha index 2 Term;  
  0, ruleabout martha index 0 Factor;  
]
```

La dérivation, représentée à l'aide d'une valeur Caml.

Derivation in Caml

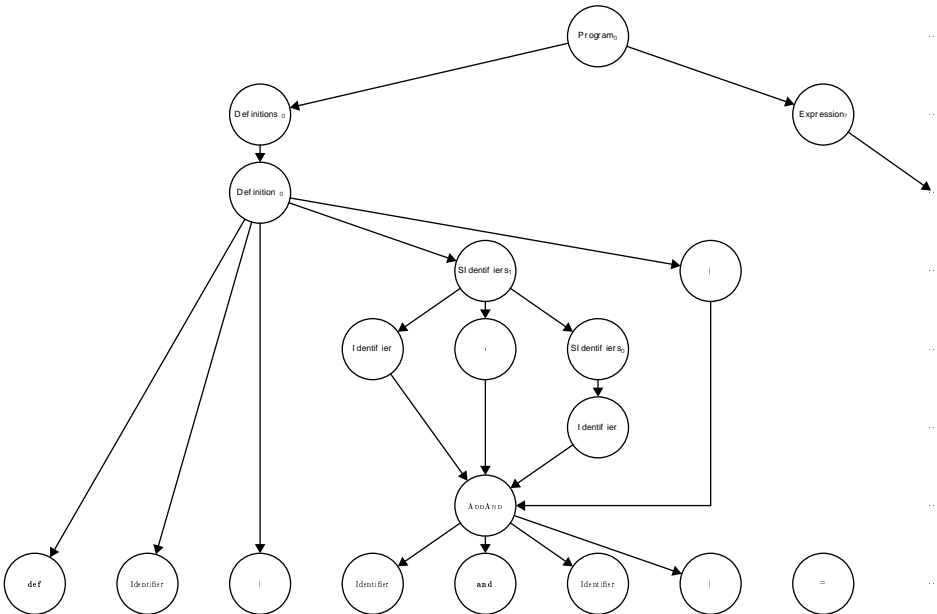
```
let apply_rule occ ((start, seq), result) s =  
  let pre x, focus = split occ in s  
  let su x =  
    let rec aux (pattern, s) =  
      match (pattern, s) with  
      | j [], su x ! -> su x  
      | j p :: ps, sym :: syms when p = sym ! -> aux (ps, syms)  
      | j _ ! -> failwith "Invalid rule application."  
    in  
    aux (start : seq, focus)  
  in  
  pre x @ result @ su x  
  
let rec interpret g s function  
  | j [] ! -> s  
  | j (occ, r) :: d ! -> interpret g (apply_rule occ r s) d
```

Représentation graphique



- Un **graphe (acyclique) de production** associé à une dérivation est formé :
 - de nœuds correspondants aux règles appliquées ;
 - d'arêtes symbolisant les entrées et sorties des règles.
- Par abus de notation, un nœud contenant un symbole correspond à la règle de reconnaissance de ce symbole.
- Dans cette représentation, les deux dérivations précédentes sont identiques.

Exemple : une règle à plusieurs entrées



Construire la représentation graphique d'une dérivation

Exercice

Instrumentez l'interprétation des dérivations que nous avons vu, afin de pouvoir, à partir de la dérivation d'une phrase, générer la représentation graphique de la dérivation.
(Vous pouvez produire un fichier dans le format d'entrée de Graphviz⁴.)

4. <http://www.graphviz.org/>

Représentations canoniques des graphes de production

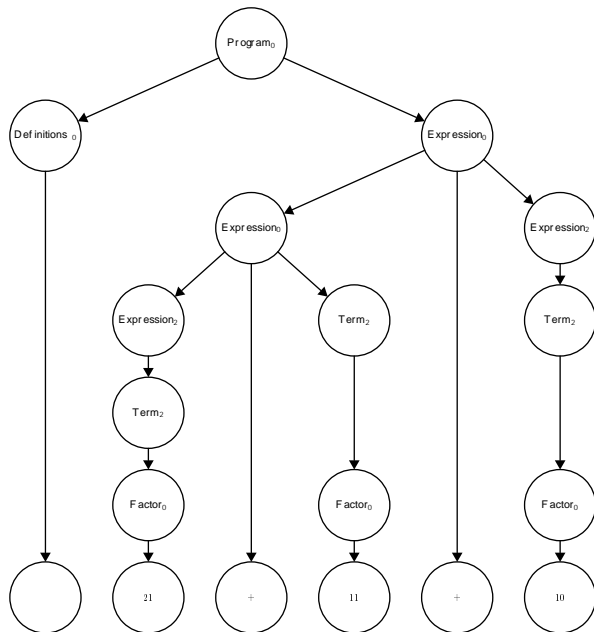
- | En effectuant un parcours ~~pré~~fixe du graphe de production, on obtient la **dérivation gauche**, c'est-à-dire la dérivation qui écrit en priorité les symboles les plus à gauche.
- | En effectuant un parcours ~~pré~~fixe, on obtient l'image miroir de la **dérivation droite**, c'est-à-dire la dérivation qui écrit en priorité les symboles les plus à droite.
-) Ces dérivations sont des représentants canoniques pour les dérivations équivalentes (qui possèdent le même arbre de production).

La grammaire du langage Martha , modi Øe

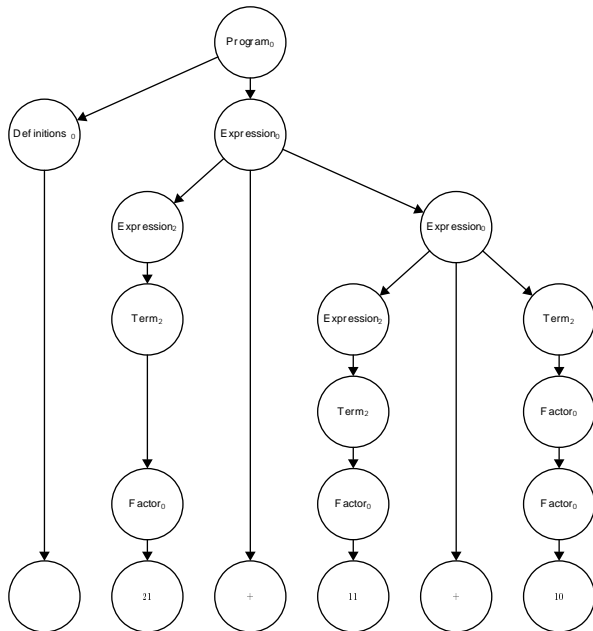
- On modifie la grammaire de Martha de la façon suivante :

	⋮	
Expression	!	Expression '+' Expression
Expression	!	Expression '-' Expression
Expression	!	Term
	⋮	

Exemple : une r  elle ambigu  t  



Exemple : une r  elle ambigu  t  



Ambiguïté

- | Une grammaire G est **ambiguë** si un mot est dérivable par deux arbres de production différents.
- | Dans l'exemple précédent, on a reconnu deux parenthésages différents de l'expression $21 + 11 + 10$: $(21 + 11) + 10$ et $21 + (11 + 10)$.
- | Ce type d'ambiguïté est bénigne dans le sens où les deux façons d'analyser l'expression conduisent à une même interprétation.
- | Cependant, certaines ambiguïtés ne le sont pas : les expressions $21 - 10$ et $(21 - 11) - 10$ valent 0 tandis que $21 - (11 - 10)$ vaut 20.

Spécification du problème de la reconnaissance

- | Le problème de l'analyse syntaxique se spécifie généralement ainsi :
 - | Entrées :
Une grammaire $G = (T; N; R; S)$, une entrée $I \in (T \cup N)^*$.
 - | Précondition :
La grammaire G est non ambiguë.
 - | Sortie :
Un arbre de production t ou bien une erreur.
 - | Postcondition :
Si la sortie est un arbre t alors il correspond à la dérivation⁵ de I par les règles de G .
Sinon, il n'existe pas de dérivation de I par les règles de G .

5. Plus précisément à un ensemble de dérivations équivalentes

Spécification du problème de la reconnaissance

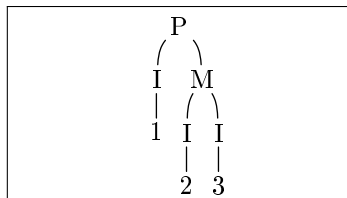
- | Le problème de l'analyse syntaxique peut aussi se spécifier ainsi :
 - | Entrées :
Une grammaire $G = (T; N; R; S)$, des règles de suppression des ambiguïtés, une entrée $I \in T^*$.
 - | Précondition :
Pas de précondition.
 - | Sortie :
Un arbre de production F ou bien une erreur.
 - | Postcondition :
Si la sortie est un arbre de production alors il existe plusieurs dérivations de I par les règles de G mais les règles de suppression des ambiguïtés ont permis d'effectuer une discrimination parmi ces dérivations non équivalentes.
Sinon, il n'existe pas de dérivation de I par les règles de G .

Exemple de règles de suppression des ambiguïtés

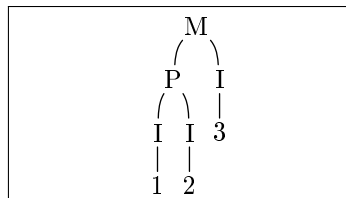
- Soit la grammaire suivante :

E	\rightarrow	i	(I)
j	\rightarrow	$E * E$	(M)
j	\rightarrow	$E + E$	(P)

- Voici deux arbres de syntaxe abstraite obtenus pour l'entrée $1 + 2 * 3$



Dérivation PIMII



Dérivation MPIII

- La règle la multiplication a une priorité plus forte que l'addition s'exprime en termes de l'arbre de production :

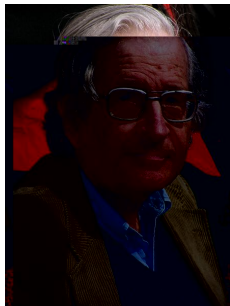
Un nœud P doit ne doit pas être le successeur direct d'un nœud M

Exemple de r gle de suppression des ambigu t s

Exercice

Comment rajouter les expressions parenth s es au langage pr c dent ?
la r gle de suppression des ambigu t s fonctionne toujours ?

Hiérarchie de Chomsky



- | Dans les années 60, Noam Chomsky a défini une **classification** des grammaires génératives s'appuyant sur des restrictions progressives de la syntaxe et des règles dans le but de faciliter leur utilisation sans pour autant restreindre de façon déraisonnable leur expressivité (leur pouvoir génératif).
- | Les grammaires de type 0 sont non restreintes.

Grammaire de type 1

- | Il y a deux définitions équivalentes des grammaires de type 1.
- | Une grammaire est de type 1 (monotone) si elle ne contient aucune règle dont le membre gauche est plus long que le membre droit.
- | Une grammaire est de type 1 (dépendante du contexte) si ses règles ne dépendent pas du contexte. Une règle est dépendante du contexte si elle est de la forme $uSv \rightarrow uvw$ où u, v et w sont des séquences de symboles (terminaux ou non terminaux) et S est un non terminal.
-) Trouver un exemple simple de langage qui ne peut pas être généré par une grammaire de type 1 est difficile.

Grammaire de type 2

- | Nous nous intéresserons surtout aux grammaires de type 2 car ce sont celles qui sont utilisées pour définir la syntaxe des langages de programmation.
- | Une grammaire est type 2 si toutes ses règles sont de la forme $A \rightarrow w$ où A est un non terminal et w est une séquence de symboles (terminaux et non terminaux).
- | C'est une grammaire de type 1 dont toutes les règles ont des contextes vides.
- | On les appelle des grammaires **hors-contexte**.
- | Dans le graphe de production, ce que produit un nœud est indépendant de ce que produisent ses voisins : le graphe de production est donc un arbre.

Grammaire de type 2

- On peut partitionner les règles d'une grammaire de type 2 suivant le non-terminal à gauche de la tête. Ce sont les règles qui dérivent ce non-terminal.
- Soit un non-terminal A . Les règles de A peuvent être regroupées en une r de la forme :

$$A \rightarrow w_1 \mid \dots \mid w_n$$

où chaque w_i est de la forme $v_1 \dots v_{k_i}$.

- Dès lors, on peut dériver le langage associé à ce non-terminal A :

$$L_A = L_{w_1} \cup \dots \cup L_{w_n}$$

où $L_{w_1} = L_{v_1} \dots L_{v_{k_1}}$, c'est-à-dire la concaténation des langages des symboles $(v_i)_{i=1 \dots k_1}$.

- Le langage L_T associé à un terminal T est bien sûr fini.

Définition récursive

- La restriction des grammaires hors-contexte n'interdit pas l'utilisation récursive d'un non-terminal dans sa définition.
- Ainsi, on peut décider si :
 - un non-terminal A est **récursif à gauche**, c'est-à-dire si il peut produire une chaîne commençant par le non-terminal A ;
 - un non-terminal A est **récursif à droite**, c'est-à-dire si il peut produire une chaîne se concluant par le non-terminal A ;
 - un non-terminal A est **auto-imbriqué**, c'est-à-dire si il peut produire une chaîne dans laquelle le terminal A se situe à gauche et à droite de deux séquences non vides.
- Les grammaires hors-contextes permettent donc d'exprimer l'**imbriquée** construction très courante des syntaxes de langage de programmation.

Notation BNF pour les grammaires hors-contextes

- | La notation BNF (Backus-Naur Form) est utilisée traditionnellement pour représenter les grammaires hors-contextes. Il en existe de très nombreuses variantes.
- | Les non-terminaux sont écrits entre chevrons : $\langle T \rangle$.
- | Les ϵ ches sont remplacées par le symbole ϵ .
- | On regroupe les règles comme indiqué plus tôt.
- | Il existe une version étendue de BNF (notée souvent EBNF) qui inclut les opérateurs $*$ (zéro ou plusieurs fois), $+$ (au moins une fois) et $?$ (plus ou moins une fois), ainsi que la possibilité de parenthéser les sous-séquences auxquelles on désire appliquer ces opérateurs.
-) Toutes ces notations sont équivalentes.

Spécification de grammaire hors-contexte

- | Dans les années 70, de nombreuses recherches ont porté sur le problème de la reconnaissance des langages décrits par les grammaires hors-contexte et leur application à la construction des compilateurs.
- | L'outil Yacc (Yet Another Compiler Compiler) a été développé à l'époque et s'appuyant sur l'algorithme d'analyse syntaxique appelé LALR(1), que nous aborderons dans ce cours.
- | Cet algorithme ne traite pas la totalité des grammaires hors-contexte. La classe des langages LALR(1) est un bon compromis entre expressivité et efficacité de l'algorithme d'analyse.
- | Nous utiliserons une version plus moderne appelée Menhir.
- | Menhir est dédié au langage OCaml et reconnaît une classe plus large de langages, les langages LR(1).

Spécification de grammaire hors-contexte

Yacc a la spécification suivante :

Entrées :

Une grammaire hors-contexte G , un ensemble de règles de suppression d'ambiguïté.

Précondition :

Pas de précondition.

Sortie :

Un **programme** qui implémente un analyseur syntaxique pour G ou bien une liste de **conits**.

Postcondition :

Si la sortie est un programme alors cela signifie que la grammaire est dans la classe des langages LALR(1) et que les règles de suppression des ambiguïtés sont complètes. Sinon, cela signifie que les règles de suppression des ambiguïtés ne sont pas suffisantes ou que la grammaire n'est pas dans la classe LALR(1).

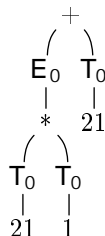
Transformations de grammaire

- | Plusieurs grammaires peuvent reconnaître le même langage.
- | Il est parfois nécessaire de **transformer** une grammaire pour la rendre compatible avec la classe de grammaires traitée par un algorithme.
- | Ces transformations nécessitent souvent l'ajout de nouveaux non-terminals et la décomposition de certaines règles.
-) On s'éloigne parfois énormément de la grammaire initiale, la plus naturelle.
- | La syntaxe d'un langage de programmation est ainsi souvent décrite par une **grammaire hors-contexte de spécification** qui définit les différentes constructions syntaxiques du langage. Cette grammaire est à destination du programmeur, puisqu'elle lui donne une vue synthétique des programmes syntaxiquement corrects. Elle sert aussi à définir la forme des arbres de syntaxe abstraits sur lesquels s'appuient les phases suivantes du compilateur.
- | L'analyse syntaxique utilise une **grammaire d'analyse** compatible avec les algorithmes d'analyse syntaxique connus et reconnaissant le même langage que la grammaire de spécification.

De l'arbre de production à l'arbre de syntaxe abstraite

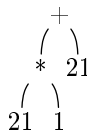
- La remarque précédente implique que l'arbre de production issu de l'analyse syntaxique fait référence aux règles de la grammaire d'analyse et est différent de l'arbre de syntaxe abstraite qui fait référence à la grammaire de spécification.
- Exemple:

E	::=	T
	j	E '+' T
T	::=	F
	j	T '*' F
F	::=	i



)?

E	::=	i
	j	E '*' E
	j	E '+' E



Action sémantique

- | En pratique, peu de systèmes produisent explicitement l'arbre de production.
- | Les outils de la famille `Yacc` exigent une **action sémantique** associée à chaque règle. Elle sert à calculer une **valeur sémantique** associée à chaque non-terminal reconnu en utilisant les valeurs sémantiques des non-terminals intervenant de la règle qui le produit.
- | En Menhir, on écrit ainsi :

```
expr: lhs=expr PLUS rhs=term
{ (* L'action sémantique est une expression OCaml. *)
  (* Elle peut faire référence à [lhs] et [rhs], *)
  (* Les valeurs sémantiques des deux sous-expressions. *)
  (* Ici, on utilise le constructeur [Add] du ty
    Add (lhs, rhs)
  }
```

Format des spécifications écrites à l'aidemenhir

```
%{
  (* Ici le prélude, c'est-à-dire du code OCaml qui peut définir *)
  (* des types de données et des fonctions utilisés dans les *)
  (* actions sémantiques. *)
}%

(* Déclarations de la grammaire. *)
%token T (* Définit un token T. *)
%token<int> I (* Définit un token I auquel est attaché un entier. *)

%start e(* Déclare le non-terminal e d'entrée de la grammaire. *)
      (* Menhir accepte plusieurs points d'entrée. *)
%type<int> e (* Déclare le type de la valeur sémantique de e. *)

(* Des priorités entre des règles (explications plus tard) *)
%left p
%right u v
%nonassoc k
%%

(* Les règles de la grammaire. *)
e : T x=e y=l { x + y }
  j T T x=e y=l { x * y }

%{
  (* Postlude *)
}%
```

Exemple de spécification écrite à l'aide de `Declar`

```
%{
  (* Abstract Syntax Tree. *)
  type exp =
    j Int of int
    j Add of exp   exp
    j Mult of exp   exp
%}

(* The lexing phase will produce the following tokens: *)
%token<int> INT
%token PLUS STAR EOF

(* Here are the declarations of non terminal symbols: *)
%type<exp> top_exp
%starttop_exp

%%

(* Now, we are defining the rules of the grammar: *)
top_exp: e=exp EOF { e }

exp: x=INT { Int x }
    j lhs=exp PLUS rhs=exp { Add (lhs, rhs) }
    j lhs=exp STAR rhs=exp { Mult (lhs, rhs) }
```


Utilisation de Menhir

- | On utilise la commande :

```
% menhir explain arith.mly
Warning 2 states have shift/reduce con icts.
Warning 4 shift/reduce con icts were arbitrarily resolved.
```

-) Menhir nous informe qu'il a bien réussi à interpréter notre grammaire en faisant des choix arbitraires pour résoudre certaines ambiguïtés.
- | Menhir a produit :
 - | un fichier `arith.ml` qui réalise l'analyse syntaxique ;
 - | un fichier `arith.ml.i` qui définit son interface ;
 - | un fichier `arith.conflicts` qui explique comment la résolution des conflits a été faite.
-) Il y a très peu de chances pour que ce soit ces choix arbitraires soient bons. Regardons ce dernier fichier !

Explication d'un con it particulier

Manhir

- ** Con ict (shift/reduce) in state 7.
- ** Tokens involved STAR PLUS
- ** The following explanations concentrate on token STAR.
- ** This state is reached from top_exp after reading

exp PLUS exp

- ** The derivations that appear below have the following common factor
- ** (The question mark symbol (?) represents the spot where the derivations begin to di er.)

top_exp
exp EOF
(?)

- ** In state 7, looking ahead at STAR, reducing production
- ** exp ! exp PLUS exp
- ** is permitted because of the following sub-derivation

exp STAR exp // lookahead token appears
exp PLUS exp .

- ** In state 7, looking ahead at STAR, shifting is permitted
- ** because of the following sub-derivation

exp PLUS exp
exp . STAR exp

Fonctionnement d'un analyseur producteur

- | Un analyseur syntaxique producteur lit le mot de lexèmes de gauche à droite.
-) Il fait partie de la famille des analyseurs **directionnels**.
- | Il essaie de reconstruire l'arbre de production en partant des feuilles.
-) C'est une analyse **ascendante**.
- | Pour cela, il maintient une pile contenant les sous-arbres de l'arbre de production déjà reconnus et prend **une** décision en fonction du sommet de cette pile et du lexème suivant.
-) L'analyseur est un **automate à pile déterministe** qui lit un lexème en a
- | Il y a deux types de décision :
 - | La décision **avance** (shift) : on pousse le lexème suivant sur la pile.
 - | La décision **réduit** (reduce) : on dépile N sous-arbres de la pile pour en construire un autre que l'on pose au sommet de la pile.

R solution d'un con it par

- | Le con it pr c dent s'explique ainsi :
 - | Apr s avoir reconnu exp PLUS exp (trois  l ments sur la pile), si le prochain lex me est STAR alors Menhir ne sait pas choisir entre :
 - | L'action avance : on empile STAR.
 - | L'action r duit : on construit une nouvelle expression correspondant au sous-arbre exp + exp que l'on met sur la pile.
 - | De toute  vidence, la seconde action n'est pas acceptable.
 - | Un autre con it appara t apr s avoir lu exp STAR exp et si le prochain lex me est PLUS . Dans ce cas, on veut plut t r duire.
-) La r gle de r duction de $*$ doit avoir une priorit  plus forte que celle de $^{00} + ^{00}$. Il su t d' crire :

(* On utilise le terminal le plus   droite de la r gle *)
(* pour parler de celle-ci. Nous verrons un mot-cl  %prec, plus  logant. *)
%nonassoc PLUS (* Les r gles du haut sont moins prioritaires ... *)
%nonassoc STAR (* ... que les r gles du bas. *)

-) 2 con its sont trait s sur 4.

R solution des con its M nhir

- | Les 2 autres con its parlent de l'interaction entre PLUS/PLUS et STAR/STAR.

Exercice

Sur quoi porte ces con its ? Quel e et a eu le mon assoc ?

Grammaire de type 3

- | Les grammaires hors-contexte décrivent des langages dans lesquels on peut-être imbriqués. Lorsque l'on analyse un sous-mot d'un mot de ces langages, la forme des règles de la grammaire nous permet de nous souvenir de ce qui doit suivre le sous-mot une fois qu'il sera reconnu.
- | La restriction des grammaires de type 3 supprime cette mémoire.
- | On formalise cette restriction de la façon suivante :
 - Une règle ne peut produire qu'un ou plusieurs terminaux suivis d'un unique non terminal optionnel.
-) En d'autres termes, les règles doivent être linéaires droites.
 - | Ces grammaires correspondent aux langages rationnels (ou réguliers) que nous avons déjà étudiés.
 - | Les automates finis sont les outils de prédilection pour la construction d'analyseurs syntaxiques de ces langages.
 - | L'expressivité des langages rationnels n'est en général pas suffisante pour décrire la syntaxe des langages de programmation. Par contre, elle suffit pour l'analyse lexicale.

Grammaire de type 4

- | La restriction des grammaires de type 4 interdit la présence d'un non terminal à droite de la tête.
-) Ces grammaires représentent des langages nis.
- | Elles fournissent un cadre pour décrire des énumérations.

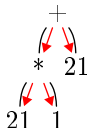
Deux directions possibles pour l'analyse syntaxique

- Rappel:

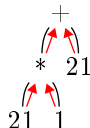
Le problème de l'analyseur syntaxique est de construire l'arbre de production d'un mot w par une grammaire non ambiguë.

- Il y a deux grandes techniques pour résoudre ce problème :

- L'analyse **descendante** essaie d'imiter la dérivation (supposée) du mot w en partant du symbole d'entrée de la grammaire G en reconstruisant l'arbre de production **de sa racine vers ses feuilles**.
- L'analyse **ascendante** tente une reconstruction de la dérivation **à partir des feuilles** de l'arbre (le mot w) en **remontant vers la racine**.



Analyse descendante



Analyse ascendante

Exemple d'analyse descendante

- Soit la grammaire (tirée de Parsing Techniques de Grune et Jacobs) :

(1)		S	!	aSQ
(2)		S	!	abc
(3)		bQc	!	bbcc
(4)		cQ	!	Qc

et l'entrée : aabbcc

⑤

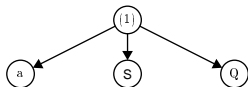
- L'analyse débute avec celle du symbole
- Quelle peut être la prochaine étape ?

Exemple d'analyse descendante

- Soit la grammaire (tirée de Parsing Techniques de Grune et Jacobs) :

(1)		S	!	aSQ
(2)		S	!	abc
(3)		bQc	!	bbcc
(4)		cQ	!	Qc

et l'entrée : aabbcc



- Comme `abc` n'est pas un préfixe de l'entrée, la règle (2) ne s'applique pas.
- Seule la règle (1) peut s'appliquer.

Exemple d'analyse descendante

- Soit la grammaire (tirée de Parsing Techniques de Grune et Jacobs) :

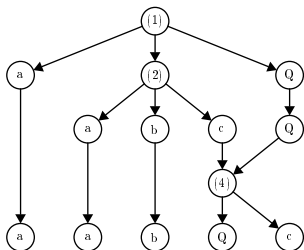
(1)		S	!	aSQ
(2)		S	!	abc
(3)				

Exemple d'analyse descendante

- Soit la grammaire (tirée de Parsing Techniques de Grune et Jacobs) :

(1)		S	!	aSQ
(2)		S	!	abc
(3)		bQc	!	bbcc
(4)		cQ	!	Qc

et l'entrée : aabbcc



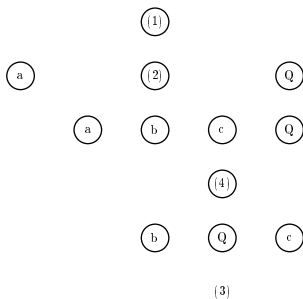
- Seule la règle (4) s'applique.
(C'était difficile à prévoir !)

Exemple d'analyse descendante

- Soit la grammaire (tirée de Parsing Techniques de Grune et Jacobs) :

(1)	S	!	aSQ
(2)	S	!	abc
(3)	bQc	!	bbcc
(4)	cQ	!	Qc

et l'entrée : aabbcc



Exemple d'analyse ascendante

- Soit la grammaire (tirée de Parsing Techniques de Grune et Jacobs) :

(1)	S	!	aSQ
(2)	S	!	abc
(3)	bQc	!	bbcc
(4)	cQ	!	Qc

et l'entrée : aabbcc

a a b b c c

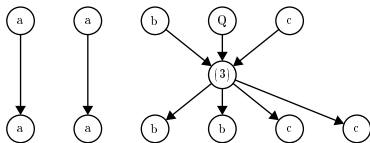
- Quelle règle peut produire une sous-séquence de non-terminaux de ce mot ?

Exemple d'analyse ascendante

- Soit la grammaire (tirée de Parsing Techniques de Grune et Jacobs) :

(1)	S	!	aSQ
(2)	S	!	abc
(3)	bQc	!	bbcc
(4)	cQ	!	Qc

et l'entrée : aabbcc



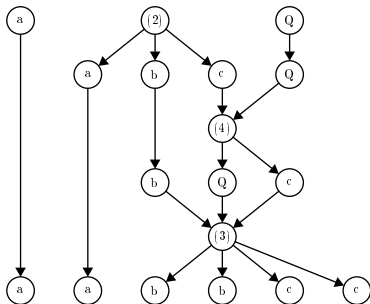
- De même, seule la règle (4) a pu produire le mot Qc .

Exemple d'analyse ascendante

- Soit la grammaire (tirée de Parsing Techniques de Grune et Jacobs) :

(1)	S	!	aSQ
(2)	S	!	abc
(3)	bQc	!	bbcc
(4)	cQ	!	Qc

et l'entrée : aabbcc



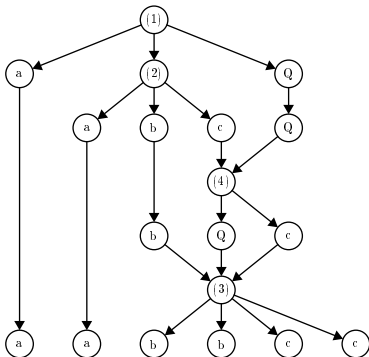
- Encore une fois, une seule règle s'applique la règle (2).

Exemple d'analyse ascendante

- Soit la grammaire (tirée de Parsing Techniques de Grune et Jacobs) :

(1)	S	!	aSQ
(2)	S	!	abc
(3)	bQc	!	bbcc
(4)	cQ	!	Qc

et l'entrée : aabbcc



- Pour finir, seule la règle (1) produit le mot aSQ .

Synthese

Grammaire et analyse syntaxique

- | Nous avons (re)vu le concept de grammaire algébrique.
- | Le problème de l'analyse syntaxique consiste à construire un arbre de production à partir d'une grammaire et d'un mot qui retrace la reconnaissance de ce mot par la grammaire, c'est-à-dire sa dérivation.
- | Menhir permet de spécifier des grammaires hors-contexte.
- | Il ne les traite pas dans leur totalité.
- | Il existe deux grands types d'analyse :

Les analyses descendantes et ascendantes

-) La définition de certains algorithmes qui les réalisent sera l'objet du prochain cours.

Bibliographie

- | Parsing techniques, A practical guide, Grune, Jacobs.
VU University Amsterdam, Amsterdam, The Netherlands
(Première Édition est disponible en ligne (<http://www.few.vu.nl/~dick/PTAPG.html>))
- | Théorie des langages, Demaille, Yvon.
Notes de cours.
(<http://www.lrde.epita.fr/~akim/thl/theorie-des-langages-1.pdf>)

QCM : Question 1

L'analyse syntaxique :

- a. produit un flux de lexème.
- b. peut être ascendante ou descendante.
- c. s'appuie sur les grammaires de type 1.
- d. peut utiliser un automate (à pile).

QCM : Question 2

On peut résoudre un conflit dans une spécification de grammaire :

- a. en modifiant le langage à reconnaître.
- b. en réécrivant la grammaire en une grammaire équivalente.
- c. en renommant des non-terminaux.
- d. en spécifiant des priorités entre les règles.

QCM : Question 3

Un graphe de production est :

- a. l'arbre de syntaxe abstraite.
- b. un graphe potentiellement cyclique.
- c. un graphe acyclique.
- d. formé de nœuds qui représentent des applications de règles de la grammaire.
- e. formé d'arêtes étiquetées par des noms de règles de la grammaire.

QCM : Question 4

L'analyse lexicale produit :

- a. un non-terminal.
- b. un flux de non-terminaux.
- c. un flux de terminaux.
- d. un flux de caractères.

QCM : Question 5

En parcourant l'arbre de production :

- a. suivant un parcours préfixe, on obtient une dérivation droite.
- b. suivant un parcours préfixe, on obtient une dérivation gauche.
- c. suivant un parcours infixe, on obtient une dérivation gauche.
- d. suivant un parcours infixe, on obtient une dérivation droite.