

Analyse Syntaxique ascendante

- OcamlYacc: un générateur d'analyseurs LALR(1)
- Structure de la source OcamlYacc
- Déclarations et typage implicite
- Utiliser les grammaires ambiguës
 - Précédence
 - Associativité
 - Traiter la Conditionnelle
- Exemples importants
 - lire la sortie de OcamlYacc
 - adapter une grammaire: opérateurs
- L'exemple classique: la calculatrice en OcamlYacc/OcamlLex.

Yacc et OcamlYacc

La construction des tables LALR(1) pour un vrai langage produit des automates avec plusieurs centaines d'états (ma grammaire pour le langage du projet en a 125), donc il faut utiliser des générateurs automatiques qui prennent une description de la grammaire et produisent un analyseur.

Tel est le but de Yacc (ou Bison) qui produisent un analyseur écrit en C, et de OcamlYacc, qui produit un analyseur écrit en Ocaml. Ces générateurs d'analyseurs sont fait pour travailler en tandem avec Lex (ou Flex) pour C, et OcamlLex, respectivement.

La grammaire des sommes, en OcamlYacc

```
La grammairea
S → E $
E → T + E | T
T → id

est décrite dans le fichier
source OcamlYacc à coté:
avue dans la partie théorique
```

```
%{
}%
/* un commentaire OcamlYacc! */
%token EOF ID PLUS
%start s
%type <unit> s
%%
s: e EOF {}
;
e: t PLUS e {}
  | t {}
;
t: ID {}
;
%%
```

Structure de la source OcamlYacc

Comme pour OcamlLex, la source est divisée en plusieurs parties logiques:

```
%{
déclarations et code Ocaml utilisés dans les actions
de l'analyseur (partie optionnelle)
}%
directives et déclarations pour OcamlYacc
%%
description des règles de la grammaire
%%
déclarations et code Ocaml qui utilisent les
fonctions d'analyses produites par OcamlYacc
(partie optionnelle)
```

Directives et déclarations pour OcamlYacc

%token *symbol...symbol* Déclaration des symboles terminaux.

%token <type> *symbol...symbol* Déclaration de tokens ayant une valeur sémantique associée de type *type*.

%start *symbol...symbol* Déclaration des points d'entrée.

%type <type> *symbol...symbol* Déclaration du type renvoyé par la fonction analysant le non terminal donné. Nécessaire seulement sur les points d'entrée (à différence de Yacc!)

Enfin, précédence et associativité (on y revient):

%left *symbol...symbol*

%right *symbol...symbol*

%nonassoc *symbol...symbol*

Description des règles de la grammaire

Les lignes

```
e: t PLUS e {}1
  | t {}2
;
```

décrivent les deux productions

$$\begin{aligned} E &\rightarrow E + T \\ E &\rightarrow T \end{aligned}$$

Entre accolades on peut mettre du code Ocaml, qui a accès à des valeurs \$1...\$n, sur lesquels on reviendra.

¹ici, des actions *sémantiques*

²ici aussi

Appel de OcamlYacc

Si `exmpl.mly` contient le source OcamlYacc, alors la commande

```
camlyacc exmpl.mly
```

produit les fichiers

exmpl.mli la description de l'interface de l'analyseur.

Cela contient aussi la définition d'un type `Ocaml token`³ contenant tous les tokens déclarés avec la directive `%token`.

exmpl.ml l'analyseur syntaxique produit par OcamlYacc.

Cela définit une fonction

```
fnt: (Lexing.lexbuf -> token) -> Lexing.lexbuf -> t
```

pour chaque symbole non terminal *fnt* déclaré dans une directive `%start` et dont le type *t* est donné dans une directive `%type`.

Appel de OcamlYacc (suite)

Si `exmpl.mly` contient le source OcamlYacc, alors la commande

```
ocaml yacc -v exmpl.mly
```

produit en plus le fichier

exmpl.output la description de l'automate LALR(1).

Utiliser des grammaires ambiguës

OcamlYacc (comme Yacc), permet (et encourage) l'utilisation de grammaires ambiguës, pour lesquelles on utilise les directives `%left`, `%right` et `%nonassoc` pour spécifier associativité et précedence:

```
/* si on veut lire -1+3*4*5=-4+3+4
   comme ((-1)+((3*4)*5))=((-4)+3)+4), on peut ecrire */
/* precedence associativity */
%nonassoc EQUAL /* faible non */
%left PLUS /* moyenne à gauche */
%left MULT /* élevée à gauche */
%nonassoc UMINUS /* plus élevée non */
```

N.B.: `nonassoc` signifie que l'opérateur n'est pas associatif, et donc en particulier `3=4=5` ne peut pas être reconnu.

³On peut ensuite utiliser ce fichier dans un fichier d'entrée pour OcamlLex, ce qui permet d'écrire la définition des tokens une seule fois à un seul endroit.

Exemple

Donc pour la grammaire

$$S \rightarrow E \$ \quad E \rightarrow E + E \mid id$$

plutôt que disambiguer⁴ on peut utiliser le source OcamlYacc suivant qui est équivalent, mais plus rapide

```
%token EOF ID PLUS
%right PLUS
%start s
%type <unit> s
%%
s:   e EOF {} ;
e:   e PLUS e {}
    | ID {} ;
```

Précédences, plus en détail

Quand OcamlYacc trouve un conflit shift/reduce dans un état, et dispose d'une série de déclarations de précédences données dans le source OcamlYacc, il procède comme suit:

- il associe à la règle utilisée pour reduce la priorité *pr* du symbole terminal le plus à droite
- il compare cette priorité avec celle *ps_t* de chaque terminal *t* pour lequel on devrait faire un shift
 - si *pr* < *ps_t* alors il choisit shift
 - si *pr* > *ps_t* alors il choisit reduce
 - si *pr* = *ps_t* alors on a déclaré les symboles sur la même ligne, avec la même directive
 - * si elle est `%left` on réduit
 - * si elle est `%right` on décale
 - * si elle est `%nonassoc` on met dans la case une action erreur

Précédences, plus en détail: exemple

Dans la grammaire ambiguë des expressions arithmétiques avec la déclaration

```
%left PLUS
%left MULT
```

on résout le conflit

S	→	E \$
E	→	T
E	→	T + E
T	→	id

```
e . MULT e
e PLUS e .
```

par un shift, parce-que MULT a précédence supérieure à PLUS, et le conflit

```
e . PLUS e
e PLUS e .
```

par un reduce, parce-que PLUS est déclaré associé à gauche

Précédences, directive %prec

On peut expliciter la précédence d'une règle avec des terminaux fictifs (on utilise ce terminal à la place du terminal le plus à droite):

$$\begin{array}{l} S \rightarrow E \$ \quad E \rightarrow id \\ E \rightarrow E + E \quad E \rightarrow \text{let } id \text{ equal } E \text{ in } E \end{array}$$

```
%token EOF ID PLUS LET EQUALS IN
%nonassoc LETPREC
%right PLUS
%start s
%type <unit> s
%%
s:   e EOF {};
e:   e PLUS e {}
    | ID {}
    | LET ID EQUALS e IN e %prec LETPREC {};
```

Précédences, if then else

Un autre cas typique est

$$\begin{array}{l} S \rightarrow E \$ \quad E \rightarrow \text{if } E \text{ then } E [\text{else } E] \\ E \rightarrow id \end{array}$$

```
%token EOF IF THEN ELSE ID
%start s
%type <unit> s
%%
s:   e EOF {};
e:   ID {}
    | IF e THEN e {}
    | IF e THEN e ELSE e {};
```

Précédences, if then else (suite)

$$\begin{array}{l} S \rightarrow E \$ \quad E \rightarrow \text{if } E \text{ then } E [\text{else } E] \\ E \rightarrow id \end{array}$$

L'ambiguïté produit le conflit:

```
10: shift/reduce conflict (shift 11, reduce 3) on ELSE
state 10
e : IF e THEN e . (3)
e : IF e THEN e . ELSE e (4)
```

```
ELSE shift 11
EOF reduce 3
THEN reduce 3
```

OcamlYacc choisit toujours *shift* dans un conflit *shift/reduce*. C'est bien ici, donc on ne modifie pas la source.

Pièges à éviter

Ceci est attrayant, ...

```
%token EOF ID PLUS MULT MINUS
%start s
%right PLUS MINUS
%left MULT
%type <unit> s
%%
s:   e EOF {};
e:   e op e {}
    | ID {};
op:
    PLUS {}
    | MULT {}
    | MINUS {};
```

Pièges à éviter

... mais mauvais (et les précédences ne règlent pas le problème)

```
11: shift/reduce conflict (shift 7, reduce 2) on PLUS
11: shift/reduce conflict (shift 8, reduce 2) on MULT
11: shift/reduce conflict (shift 9, reduce 2) on MINUS
state 11
e : e . op e (2)
e : e op e . (2)
```

```
PLUS shift 7
MULT shift 8
MINUS shift 9
EOF reduce 2
```

```
op goto 10
```

Pièges à éviter

Il faut

```
%token EOF ID PLUS MULT MINUS
%start s
%right PLUS MINUS
%left MULT
%type <unit> s
%%
s: e EOF {};
e:
    e PLUS e {}
  | e MULT e {}
  | e MINUS e {}
  | ID {};
```