
Rapport sur le transpilateur MiniML vers C

Name
22 février 2018

TABLE DES MATIÈRES

1	Analyses	2
2	Analyse lexicale	3
2.1	Tokens et lexèmes	3
2.2	Commentaires	3
3	Analyse syntaxique	4
3.1	Structure d'AST	4
3.2	Bison et construction de l'AST	5
4	Annexe 1 : Grammaire	6
4.1	Grammaire BNF	6
4.2	Terminaux	7

1 ANALYSES

La compilation est divisée en quatre étapes :

1. l'analyse lexicale (transformation du texte en flot de lexèmes)
2. l'analyse syntaxique (transformation du flot de lexèmes en arbre syntaxique abstrait)
3. l'analyse sémantique (vérification de la validité des expressions)
4. l'exécution ou la génération du code

Chaque section traitera un type d'analyse.

NB : Le programme final doit être compilé et lié avec libcalg ([à compiler à partir des sources](#)). Il s'agit d'une bibliothèque de structures de données, utilisée dans le but de simplifier les implémentations, i.e. le code relatif aux structures des données de base est abstrait par libcalg.

2 ANALYSE LEXICALE

2.1 TOKENS ET LEXÈMES

L'implémentation présente permet 6 types de tokens différents (association d'un lexème avec une valeur éventuelle).

LES LITÉRAUX : Les nombres entiers, les nombres flottants et les chaînes de caractères ont leur propre catégorie. Les lexèmes `true` et `false` sont associés dans une même catégorie aux valeurs 1 et 0 respectivement.

LES IDENTIFIANTS : Les identifiants et les opérateurs appartiennent à la même catégorie des noms. En effet, par la nature fonctionnelle des langages ML, les opérateurs sont considérés comme des fonctions infixes, et donc des identifiants spéciaux.

Une table des noms est mise en place pour associer chaque identifiant à un entier unique afin de simplifier la comparaison d'identifiants. Cette table est implémentée par une table d'association bijective.

Chaque nouveau nom est associé à un entier unique strictement supérieur à 0. L'entier 0 est utilisé lorsqu'un identifiant ne peut pas exister (p.ex. une fonction anonyme).

LES AUTRES : Les autres mots-clés et lexèmes restants ne portant pas de valeur particulière, ils forment la dernière catégorie de tokens. Les caractères blancs sont ignorés par l'analyseur, et les caractères invalides sont passés à l'analyseur syntaxique tels quels, où ils provoqueront une erreur de syntaxe.

2.2 COMMENTAIRES

L'analyseur lexical ignore les commentaires, lesquels peuvent être imbriqués.

Le contexte devient `comment` en lisant un commentaire. Un compteur est incrémenté et décrementé en cas d'ouverture et de fermeture de commentaire respectivement. Quand le compteur est nul, le contexte redevient le contexte initial.

Les spécificités dans l'implémentation de cette règle permettent d'optimiser la lecture quand les caractères des balises de commentaire apparaissent dans son contenu.

REMARQUE : Il peut être plus efficace de créer une fonction `C` pour analyser les commentaires. En effet, en pratique, le tampon de lecture de Flex a une capacité limitée, et si la longueur du commentaire dépasse cette capacité, le comportement du programme peut être indéterminé.

3 ANALYSE SYNTAXIQUE

3.1 STRUCTURE D'AST

La structure d'AST utilisée contient un champ d'énumération de type et une union anonyme selon ce champ.

Type	Champs
unit	-
integer	value: Int
float	value: Float
boolean	value: Bool
string	value: String
variable	name: Int
block	expr: AST
list	list: List[AST]
application	function: AST, args: List[AST]
let	names: List[AST], rec: Bool, params: List[AST], expr: AST, block: AST
if	cond: AST, if: AST, else: AST
tuple	list: List[AST]

Voici une explication détaillée pour la construction let :

Le champ names est une liste des noms liés par cette expression. Cette liste est vide si et seulement si l'expression est une déclaration de fonction anonyme avec function. La liste contient plus d'un nom dans le cas où l'expression définit un tuple de noms.

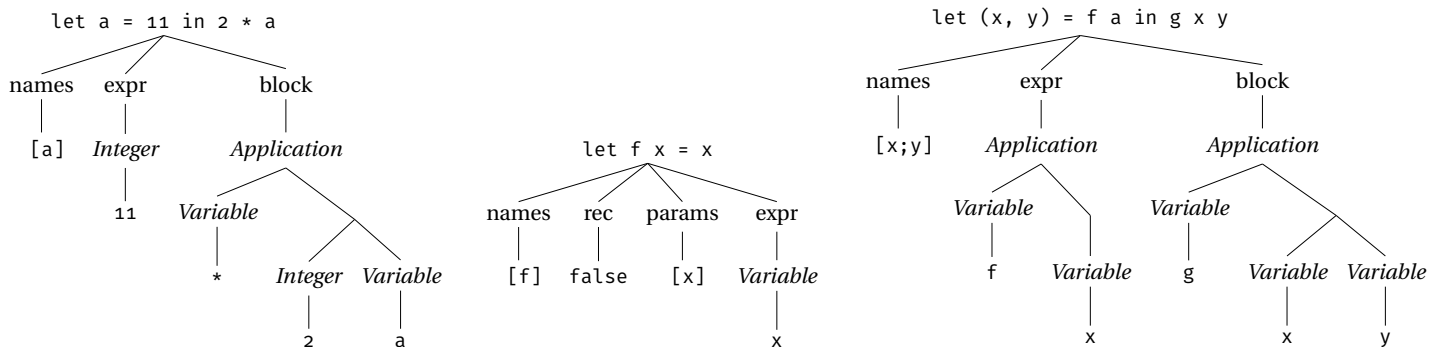
Le champ rec est une valeur booléenne. Elle n'a de sens qu'en déclaration de fonction. Le champ vaut true si et seulement si la définition est récursive.

Le champ params est la liste des noms des paramètres de la fonction déclarée. Cette liste est vide si et seulement si il s'agit d'une déclaration de variable.

Le champ expr est l'expression du corps de la déclaration.

Le champ block est l'expression suivant cette définition. Il correspond au bloc in dans la construction let-in.

Et quelques exemples :



On remarquera que l'opération `2 * a` est équivalente à l'application `(*) 2 a`, comme décrit avant.

3.2 BISON ET CONSTRUCTION DE L'AST

La grammaire du langage est décrite en annexe dans le méta-langage BNF. Les modifications suivantes ont été effectuées pour l'adapter au générateur Bison.

La récupération sur erreur s'effectue par instruction : en cas d'erreur de syntaxe, l'instruction en cours d'analyse est ignorée et celle-ci reprend à la fin de l'instruction erronée.

La grammaire décrite contient également des conflits décalage/réduction. Ceux-ci ont été résolus en appliquant des règles de priorité et d'associativité sur certaines productions et terminaux.

Les priorités des opérations arithmétiques sont définies explicitement, en créant des règles intermédiaires pour former `simple-expr`, en suivant l'ordre conventionnel des opérations.

Voici, pour illustrer, la réduction de l'expression $1 = 2 + 3 * 6$:

$$\begin{aligned}
 \langle expr \rangle &\Rightarrow \langle or \rangle \Rightarrow \langle and \rangle \Rightarrow \langle equ \rangle \\
 &\Rightarrow \langle rel \rangle = \langle rel \rangle \\
 &\Rightarrow \langle add \rangle = \langle add \rangle \\
 &\Rightarrow \langle mul \rangle = (\langle add \rangle + \langle mul \rangle) \\
 &\Rightarrow \langle un \rangle = (\langle mul \rangle + (\langle un \rangle * \langle un \rangle)) \\
 &\Rightarrow 1 = (\langle un \rangle + (3 * 6)) \\
 &\Rightarrow 1 = (2 + (3 * 6)).
 \end{aligned}$$

La priorité des non-opérateurs est définie dans le fichier source Bison, grâce aux instructions de priorité/associativité.

Le tableau de priorité/associativité est le suivant, par ordre de priorité de la plus élevée à la plus faible :

Production ou symbole	Associativité
!	droite
application	gauche
+ - +. -. (préfixe)	droite
* / *. /.	gauche
+ - +. -.	gauche
= <> < <= > >=	gauche
&&	droite
	droite
,	-
else	droite
if	-
;	droite
let function	-

4 ANNEXE 1 : GRAMMAIRE

4.1 GRAMMAIRE BNF

$\langle \text{program} \rangle$	$::= \langle \text{program} \rangle \langle \text{instruction} \rangle \text{';;'}$ $ \langle \text{program} \rangle \text{';;'}$ $ \langle \text{instruction} \rangle \text{';;'}$ $ \text{';;'}$
$\langle \text{instruction} \rangle$	$::= \langle \text{expr} \rangle \mid \langle \text{let-binding} \rangle$
$\langle \text{expr} \rangle$	$::= \langle \text{simple-expr} \rangle$ $ \langle \text{atom} \rangle \langle \text{atom-list} \rangle$ $ \langle \text{let-binding} \rangle \text{'in'} \langle \text{expr} \rangle$ $ \text{'function'} \langle \text{parameter-list} \rangle \text{'->'} \langle \text{expr} \rangle$ $ \langle \text{if-expr} \rangle$
$\langle \text{simple-expr} \rangle$	$::= \langle \text{simple-expr} \rangle \text{' '} \langle \text{simple-expr} \rangle$ $ \langle \text{simple-expr} \rangle \text{'\&\&'} \langle \text{simple-expr} \rangle$ $ \langle \text{simple-expr} \rangle \langle \text{equ-op} \rangle \langle \text{simple-expr} \rangle$ $ \langle \text{simple-expr} \rangle \langle \text{rel-op} \rangle \langle \text{simple-expr} \rangle$ $ \langle \text{simple-expr} \rangle \langle \text{add-op} \rangle \langle \text{simple-expr} \rangle$ $ \langle \text{simple-expr} \rangle \langle \text{mul-op} \rangle \langle \text{simple-expr} \rangle$ $ \langle \text{unary-op} \rangle \langle \text{simple-expr} \rangle$ $ \langle \text{simple-expr} \rangle$
$\langle \text{if-expr} \rangle$	$::= \text{'if'} \langle \text{expr} \rangle \text{'then'} \langle \text{expr} \rangle$ $ \text{'if'} \langle \text{expr} \rangle \text{'then'} \langle \text{expr} \rangle \text{'else'} \langle \text{expr} \rangle$
$\langle \text{let-binding} \rangle$	$::= \text{'let'} \langle \text{let-pattern} \rangle \text{'='} \langle \text{expr} \rangle$ $ \text{'let'} \text{'rec'} \langle \text{let-pattern} \rangle \text{'='} \langle \text{expr} \rangle$
$\langle \text{let-pattern} \rangle$	$::= \langle \text{name} \rangle$ $ \text{'('} \langle \text{operator} \rangle \text{'}'}$ $ \text{'('} \langle \text{tuple-name-list} \rangle \text{'}'}$
$\langle \text{atom} \rangle$	$::= \langle \text{integer} \rangle \mid \langle \text{float} \rangle$ $ \langle \text{boolean} \rangle \mid \langle \text{string} \rangle$ $ \langle \text{name} \rangle$ $ \text{'('} \text{'}'}$ $ \text{'('} \langle \text{operator} \rangle \text{'}'}$ $ \text{'('} \langle \text{expr-list} \rangle \text{'}'}$ $ \text{'begin'} \text{'end'}$ $ \text{'begin'} \langle \text{expr-list} \rangle \text{'end'}$ $ \text{'('} \langle \text{tuple-expr-list} \rangle \text{'}'}$
$\langle \text{atom-list} \rangle$	$::= \langle \text{atom} \rangle$ $ \langle \text{atom-list} \rangle \langle \text{atom} \rangle$
$\langle \text{expr-list} \rangle$	$::= \langle \text{expr} \rangle$ $ \langle \text{expr-list} \rangle \langle \text{expr} \rangle$

$\langle parameter-list \rangle$	$::= \langle name \rangle$ $\langle parameter-list \rangle \langle name \rangle$
$\langle tuple-name-list \rangle$	$::= \langle name \rangle$ $\langle tuple-name-list \rangle \text{ ',' } \langle name \rangle$
$\langle tuple-expr-list \rangle$	$::= \langle expr \rangle$ $\langle tuple-expr-list \rangle \text{ ',' } \langle expr \rangle$
$\langle equ-op \rangle$	$::= \text{'='} \text{ ' <' }$
$\langle rel-op \rangle$	$::= \text{'>' '>=' '<' '<='}$
$\langle add-op \rangle$	$::= \text{'+' '-' '+.' '-.'}$
$\langle mul-op \rangle$	$::= \text{'*' '/' '*.' '/.'}$
$\langle unary-op \rangle$	$::= \text{'!' '+' '-' '+.' '-.'}$

4.2 TERMINAUX

INTEGER : les nombres entiers.

FLOAT : les nombres flottants, avec décimales et/ou exposant. (ex : 5., 3.1415, 9e+5 ou 1.2e-67)

BOOLEAN : les valeurs de vérité true et false.

STRING : les chaînes littérales, avec les échappements suivants : le guillemet (\"), l'apostrophe (\'), le retour à la ligne (\n), le retour arrière (\b), la tabulation (\t), le retour chariot (\r), l'espace (\).

NAME : les identifiants, commencent par une lettre et constitués de lettres, chiffres et/ou tiret du bas.

OPERATOR : les terminaux des opérateurs, â, || et les autres terminaux qui apparaissent dans les productions *-op.