École Normale Supérieure

Langages de programmation et compilation

examen 2019–2020

Jean-Christophe Filliâtre

24 janvier 2020

Les notes de cours manuscrites ou reprographiées sont les seuls documents autorisés.
L'épreuve dure 3 heures.

Dans tout ce sujet, on considère un petit langage de programmation qui est une variante du langage mini-ML étudié en cours. La syntaxe abstraite de ce langage est donnée figure 1. Cette syntaxe distingue des expressions déjà évaluées, dites *atomiques* et notées a, et des expressions arbitraires représentant des calculs (potentiels), notées e. Dans ce langage, les valeurs sont réduites à des entiers relatifs, notés n, et des fonctions récursives. Une fonction récursive est notée `rec` f x `=` e où f est une variable dénotant la fonction et x une variable dénotant son argument.

Une sémantique opérationnelle à petits pas est donnée pour ce langage, figure 2, sous la forme d'un jugement $e \to e'$ défini par des règles d'inférence.

Question 1 Donner la suite des réductions pour chacune des trois expressions suivantes :

1. `let double = rec _ n = add n n in double 21`
2. `let f = rec f n = f n in f 0`
3. `let f = rec _ n = ifz x then 34 else 55 in let x = 0 in f x`

Question 2 Donner une expression e telle que, pour tout entier naturel n, l'évaluation de l'expression `let` $f = e$ `in` $f\ n$ aboutisse à la valeur du n-ième terme de la suite de Fibonacci, c'est-à-dire `let` $f = e$ `in` $f\ n \to^\star F_n$ avec F_n ainsi défini :

$$\left\{ \begin{array}{lcl} F_0 & = & 0 \\ F_1 & = & 1 \\ F_{n+2} & = & F_n + F_{n+1} \quad \text{pour } n \geq 0. \end{array} \right.$$

La complexité de cette évaluation, définie comme le nombre total de petits pas, doit être $O(n)$.

Analyse syntaxique. On souhaite réaliser l'analyse syntaxique de notre petit langage avec l'outil Menhir. La figure 3 contient un fichier d'entrée pour Menhir contenant la grammaire de notre langage. Les actions sémantiques ne nous intéressent pas ici et sont omises (`{...}`).

Question 3 Donner les valeurs de NULL, FIRST et FOLLOW pour les deux non-terminaux `expr` et `atom`.

Question 4 Lorsque l'outil Menhir est lancé sur le fichier donné en figure 3, il déclare trois conflits de type lecture/réduction (`shift/reduce`). Les identifier, les expliquer, faire le choix de favoriser lecture ou réduction et modifier en conséquence le fichier Menhir.

Sucre syntaxique. Telle quelle, la grammaire de notre langage n'est pas très agréable, surtout lorsqu'on souhaite définir ou utiliser des fonctions à plusieurs arguments. Ainsi, pour définir une fonction à deux arguments, on doit écrire `rec` f x_1 `=` `rec` `_` x_2 `=` e et pour l'appliquer à deux arguments on doit écrire `let` $g = f\ a_1$ `in` $g\ a_2$.

Question 5 Proposer une modification de la syntaxe *concrète* pour autoriser une définition de fonction de la forme `rec` $f\ x_1\ \ldots\ x_n$ `=` e, c'est-à-dire avec $n \geq 1$ paramètres formels, ainsi qu'une application de la forme $a_0\ a_1\ a_2\ \ldots\ a_n$, c'est-à-dire avec $n \geq 1$ paramètres effectifs. Indiquer les modifications à apporter à la grammaire Menhir, ainsi que les actions sémantiques des règles qui ont été modifiées. On suppose que la syntaxe abstraite OCaml est de la forme suivante :

```
type atom = Var of string | Rec of string * string * expr | ...
and  expr = Atom of atom | App of atom * atom | Let of string * expr * expr | ...
```

On prendra soin de construire des arbres de syntaxe abstraite bien typés, en respectant la différence entre les types `atom` et `expr`.

Compilation vers x86-64. On se propose maintenant de compiler notre petit langage vers l'assembleur x86-64. (Un aide-mémoire est donné en annexe.) La compilation d'une expression e est un code assembleur noté $C(e)$ dont l'exécution a pour effet de stocker la valeur de l'expression e dans le registre `%rax`.

Notre langage étant un langage fonctionnel par nature, sa compilation met en œuvre des fermetures, comme expliqué en cours. On adopte le schéma de compilation suivant. Une valeur est soit un entier 64 bits signé, soit un pointeur vers un bloc alloué sur le tas contenant une fermeture. Les valeurs ont donc toutes la même taille (64 bits). Pour toute expression `rec` $f\ x$ `=` e contenue dans le programme, on introduit une fonction assembleur f, recevant l'argument x dans le registre `%rdi` et la fermeture dans le registre `%rsi`. Le tableau d'activation de la fonction a la forme ci-contre, les adresses croissant vers le haut. Dans ce tableau, on trouve les m variables locales à l'expression e. Les variables libres de e, en revanche, se trouvent dans la fermeture, à l'exception de la variable x qui se trouve dans `%rdi`. On note que si la fonction f est effectivement récursive, f apparaît dans les variables libres de e et peut donc être retrouvée dans la fermeture, comme toute autre variable libre de e. Le résultat de la fonction (c'est-à-dire la valeur de l'expression e constituant le corps de la fonction) est placé dans le registre `%rax`.

	adresse retour
`%rbp` →	`%rbp` sauvegardé
	variable locale 1
	⋮
	variable locale m

Question 6 On considère l'exécution du code assembleur obtenu pour le programme suivant :

```
let f = rec f x = ifz x then 42 else let x = add x -1 in f x in f 1729
```

Indiquer combien de fermetures seront allouées sur le tas pendant cette exécution.

Question 7 Donner un code assembleur pour la fonction `f` résultant de la compilation de l'expression

```
rec f x = ifz x then 42 else let x = add x -1 in f x
```

Indiquer si l'appel à `f` est terminal. Le cas échéant, l'optimiser.

Question 8 Donner la définition de $C(e)$ pour chacune des constructions du langage.

$$
\begin{array}{rcll}
a & ::= & x & \textit{variable} \\
& | & n & \textit{constante } (n \in \mathbb{Z}) \\
& | & \texttt{rec } x\ x\ \texttt{=}\ e & \textit{fonction récursive} \\
\\
e & ::= & a & \textit{atome} \\
& | & a\ a & \textit{application} \\
& | & \texttt{add } a\ a & \textit{addition} \\
& | & \texttt{let } x = e\ \texttt{in } e & \textit{variable locale} \\
& | & \texttt{ifz } a\ \texttt{then } e\ \texttt{else } e & \textit{conditionnelle}
\end{array}
$$

FIGURE 1 – Syntaxe abstraite.

$$
\frac{}{(\texttt{rec } f\ x\ \texttt{=}\ e)\ a \to e[x \leftarrow a][f \leftarrow \texttt{rec } f\ x\ \texttt{=}\ e]} \qquad \frac{n = n_1 + n_2}{\texttt{add } n_1\ n_2 \to n}
$$

$$
\frac{}{\texttt{let } x = a\ \texttt{in } e_2 \to e_2[x \leftarrow a]} \qquad \frac{e_1 \to e_1'}{\texttt{let } x = e_1\ \texttt{in } e_2 \to \texttt{let } x = e_1'\ \texttt{in } e_2}
$$

$$
\frac{}{\texttt{ifz } 0\ \texttt{then } e_1\ \texttt{else } e_2 \to e_1} \qquad \frac{n \neq 0}{\texttt{ifz } n\ \texttt{then } e_1\ \texttt{else } e_2 \to e_2}
$$

FIGURE 2 – Sémantique opérationnelle à petits pas.

```
%token <string> IDENT
%token <int> CONST
%token LET IN REC IFZ THEN ELSE ADD EQUAL EOF
%start one_expr
%type  <unit> one_expr
%%
one_expr:
  | expr EOF                      {...}
expr:
  | atom                          {...}
  | atom atom                     {...}
  | ADD atom atom                 {...}
  | LET IDENT EQUAL expr IN expr  {...}
  | IFZ atom THEN expr ELSE expr  {...}
atom:
  | IDENT                         {...}
  | CONST                         {...}
  | REC IDENT IDENT EQUAL expr    {...}
```

FIGURE 3 – Fichier Menhir.

$$
\begin{array}{rcll}
e & ::= & \ldots & \\
 & | & \texttt{let}\ x = \texttt{do}\ E\ a\ \texttt{in}\ e & \textit{déclenchement} \\
 & | & \texttt{handle}\ e\ \texttt{with}\ h & \textit{traitement} \\
h & ::= & \{\, E\ x\ x\ \texttt{=>}\ e\texttt{;}\ \ldots\texttt{;}\, E\ x\ x\ \texttt{=>}\ e\,\} & \textit{gestionnaire}
\end{array}
$$

FIGURE 4 – Syntaxe abstraite augmentée.

$$
\frac{}{\texttt{handle}\ a\ \texttt{with}\ h \to a} \qquad \frac{e_1 \to e_2}{\texttt{handle}\ e_1\ \texttt{with}\ h \to \texttt{handle}\ e_2\ \texttt{with}\ h}
$$

$$
\frac{\text{(on suppose que } y \text{ n'est pas libre dans } e_2\text{, quitte à la renommer)}}{\texttt{let}\ x = \texttt{let}\ y = \texttt{do}\ E\ a\ \texttt{in}\ e_1\ \texttt{in}\ e_2 \to \texttt{let}\ y = \texttt{do}\ E\ a\ \texttt{in let}\ x = e_1\ \texttt{in}\ e_2}
$$

$$
\frac{E \notin h}{\texttt{handle let}\ y = \texttt{do}\ E\ a\ \texttt{in}\ e_2\ \texttt{with}\ h \to \texttt{let}\ y = \texttt{do}\ E\ a\ \texttt{in handle}\ e_2\ \texttt{with}\ h}
$$

$$
\frac{(E\ x\ k\ \texttt{=>}\ e_1) \in h}{\texttt{handle let}\ y = \texttt{do}\ E\ a\ \texttt{in}\ e_2\ \texttt{with}\ h \to e_1[x \leftarrow a][k \leftarrow \texttt{rec}\ \_\ y\ \texttt{=}\ \texttt{handle}\ e_2\ \texttt{with}\ h]}
$$

FIGURE 5 – Sémantique opérationnelle à petits pas (nouvelles règles).

Effets algébriques. On augmente maintenant notre petit langage avec de nouvelles constructions, appelées *effets algébriques*, dont la syntaxe est donnée figure 4. Un effet, noté E, est un simple identifiant. La construction `let` x `=` `do` $E\ a$ `in` e déclenche l'effet E, en lui associant la valeur a, et lie la variable x dans l'expression e. La construction `handle` e `with` h évalue l'expression e, en traitant les effets résultants, le cas échéant, à l'aide du gestionnaire h. Ce dernier contient des quadruplets de la forme $E\ v\ k$ `=>` e', où les variables v et k sont liées dans l'expression e'.

Des règles de sémantique opérationnelle pour ces nouvelles constructions sont données figure 5, où la notation $E \notin h$ signifie qu'il n'existe aucun quadruplet pour E dans le gestionnaire h. Ces nouvelles règles s'ajoutent aux règles précédentes (de la figure 2), qui sont inchangées. La sémantique de ces deux constructions est assez subtile et on prendra le temps de bien lire et de bien comprendre ces nouvelles constructions. En particulier, on notera comment, dans un gestionnaire $E\ x\ k$ `=>` e, la fonction k permet de *reprendre* le calcul à l'endroit où l'effet a été déclenché, la valeur passée en argument à k se retrouvant liée à la variable y introduite par la construction `let y = do` à l'endroit du déclenchement.

Question 9 Donner la suite des réductions pour chacune des deux expressions suivantes :

1. `handle let x = do Stop 41 in x with { Stop _ k => k 42 }`
2. `handle let x = do E 21 in add x x with { E n k => let n = k n in k n }`

$$\frac{x : \tau \in \Gamma}{\Gamma \vdash x : \tau} \qquad \frac{}{\Gamma \vdash n : \texttt{int}} \qquad \frac{\Gamma, f : \tau \ \to \ \kappa, x : \tau \vdash e : \kappa}{\Gamma \vdash \texttt{rec}\ f\ x = e : \tau \ \to \ \kappa}$$

$$\frac{\Gamma \vdash a : \tau}{\Gamma \vdash a : \tau\langle S\rangle} \qquad \frac{\Gamma \vdash a_1 : \tau \ \to \ \kappa \quad \Gamma \vdash a_2 : \tau}{\Gamma \vdash a_1\ a_2 : \kappa} \qquad \frac{\Gamma \vdash a_1 : \texttt{int} \quad \Gamma \vdash a_2 : \texttt{int}}{\Gamma \vdash \texttt{add}\ a_1\ a_2 : \texttt{int}\langle S\rangle}$$

$$\frac{\Gamma \vdash e_1 : \tau_1\langle S\rangle \quad \Gamma, x : \tau_1 \vdash e_2 : \tau_2\langle S\rangle}{\Gamma \vdash \texttt{let}\ x = e_1\ \texttt{in}\ e_2 : \tau_2\langle S\rangle} \qquad \frac{\Gamma \vdash a : \texttt{int} \quad \Gamma \vdash e_1 : \kappa \quad \Gamma \vdash e_2 : \kappa}{\Gamma \vdash \texttt{ifz}\ a\ \texttt{then}\ e_1\ \texttt{else}\ e_2 : \kappa}$$

$$\frac{\Gamma \vdash a : \texttt{int} \quad \Gamma, x : \texttt{int} \vdash e : \tau\langle S\rangle \quad E \in S}{\Gamma \vdash \texttt{let}\ x = \texttt{do}\ E\ a\ \texttt{in}\ e : \tau\langle S\rangle}$$

$$\frac{\Gamma \vdash e : \tau\langle S\rangle \quad \forall i. \Gamma, x_i^1 : \texttt{int}, x_i^2 : \texttt{int} \ \to \ \tau\langle S'\rangle \vdash e_i : \tau\langle S'\rangle \quad S \backslash \{\, E_1, \ldots, E_n \,\} \subseteq S'}{\Gamma \vdash \texttt{handle}\ e\ \texttt{with}\ \{\, E_1\ x_1^1\ x_1^2\ \texttt{=>}\ e_1\texttt{;}\ \ldots\ \texttt{;} E_n\ x_n^1\ x_n^2\ \texttt{=>}\ e_n \,\} : \tau\langle S'\rangle}$$

FIGURE 6 – Règles de typage.

Question 10 Compléter l'expression suivante

```
handle
   let _ = do Write n1 in let x = do Read 0 in let x = add x n2 in
   let _ = do Write x in let y = do Read 0 in y
with { ... }
```

avec un gestionnaire (la partie indiquée `...`) de telle façon que l'évaluation de l'expression donne toujours l'entier $n_1 + n_2$. Le gestionnaire ne doit pas dépendre des constantes n_1 et n_2.

Question 11 Expliquer comment encoder la notion usuelle d'exceptions à l'aide d'effets algébriques.

Typage. On s'intéresse maintenant au typage de notre langage, avec pour objectif d'obtenir la sûreté du typage, *i.e.*, l'évaluation d'une expression bien typée aboutit à une valeur (constante entière ou fonction) ou ne termine pas. En particulier, elle ne doit pas aboutir à un déclenchement d'effet qui ne serait pas traité par un gestionnaire.

On se donne des types τ pour les valeurs et des types κ pour les calculs, définis de la manière suivante :

$$\begin{array}{rcll} \tau & ::= & \texttt{int} & \textit{type d'un entier} \\ & | & \tau \ \to \ \kappa & \textit{type d'une fonction} \\ \kappa & ::= & \tau\langle S\rangle & \textit{type d'un calcul} \\ S & ::= & \{\, E, \ldots, E \,\} & \textit{ensemble fini d'effets} \end{array}$$

Un type de calcul $\tau\langle S\rangle$ est une paire formée d'un type de valeur (le type de la valeur renvoyée par ce calcul) et d'un ensemble d'effets *susceptibles* d'être déclenchés par ce calcul. Un environnement de typage Γ donne un type de valeur τ à toute variable apparaissant dans le programme. On a deux jugements de typage : un jugement $\Gamma \vdash a : \tau$ pour les atomes et un jugement $\Gamma \vdash e : \kappa$ pour les expressions. Les règles de typage sont données figure 6. Par simplicité, on suppose que tous les effets attendent un argument de type `int` et "renvoient" une valeur de type `int`.

Question 12 Pour chacune des trois expressions suivantes, indiquer si elle est bien typée (dans un environnement vide) :

1. `let x = do E 42 in x`
2. `rec f x = f x`
3. `handle 1 with { A x k => let _ = do B 2 in 3 }`

Le cas échéant, donner un type possible et une dérivation de typage. Dans le cas contraire, justifier.

Question 13 Montrer la propriété de progrès : pour une expression e *close*, si $\emptyset \vdash e : \tau\langle\emptyset\rangle$, alors e est une valeur (un atome qui n'est pas une variable) ou il existe une expression e' telle que $e \to e'$.

Question 14 Montrer la propriété de préservation du typage : si $\Gamma \vdash e : \kappa$ et $e \to e'$ alors $\Gamma \vdash e' : \kappa$. On admettra la propriété de préservation du typage par substitution : si $\Gamma, x : \tau \vdash e : \kappa$ et $\Gamma \vdash a : \tau$, alors $\Gamma \vdash e[x \leftarrow a] : \kappa$.

Question 15 En déduire la propriété de sûreté du typage, que l'on énoncera précisément.

Compilation. On se pose enfin la question de la compilation de notre langage avec effets algébriques. On se propose de le traduire vers le langage OCaml. L'idée est d'utiliser une exception OCaml pour réaliser le déclenchement d'un effet. Cette exception prend en argument le nom de l'effet (de type `string`), la valeur passée en argument à l'effet, ainsi que la suite du calcul sous la forme d'une fonction :

```
exception Do of string * int * (int -> int)
```

Pour simplifier les choses, on suppose ici que la suite du calcul (l'expression e dans la construction `let` $x =$ `do` E a `in` e) est toujours de type `int`.

Question 16 Donner le schéma de la traduction des constructions `let` $x =$ `do` E a `in` e et `handle` e `with` h.

Annexe : aide-mémoire x86-64

On donne ici un fragment du jeu d'instructions x86-64. Vous êtes libre d'utiliser tout autre élément de l'assembleur x86-64. Dans ce qui suit, r_i désigne un registre, n une constante entière et L une étiquette.

`mov` r_2`,` r_1	copie le registre r_2 dans le registre r_1
`mov` \$$n$`,` r_1	charge la constante n dans le registre r_1
`mov` L`,` r_1	charge la valeur à l'adresse L dans le registre r_1
`mov` \$$L$`,` r_1	charge l'adresse de l'étiquette L dans le registre r_1
`add` r_2`,` r_1	calcule la somme de r_1 et r_2 dans r_1 (on a de même `sub` et `imul`)
`mov` n`(`r_2`),` r_1	charge dans r_1 la valeur contenue en mémoire à l'adresse $r_2 + n$
`mov` r_1`,` n`(`r_2`)`	écrit en mémoire à l'adresse $r_2 + n$ la valeur contenue dans r_1
`push` r_1	empile la valeur contenue dans r_1
`pop` r_1	dépile une valeur dans le registre r_1
`cmp` r_2`,` r_1	positionne les drapeaux en fonction de la valeur de $r_1 - r_2$
`test` r_2`,` r_1	positionne les drapeaux en fonction de la valeur de $r_1 \& r_2$
`je` L	saute à l'adresse désignée par l'étiquette L en cas d'égalité (on a de même `jne`, `jg`, `jge`, `jl` et `jle`)
`jmp` L	saute à l'adresse désignée par l'étiquette L
`call` L	saute à l'adresse désignée par l'étiquette L, après avoir empilé l'adresse de retour
`ret`	dépile une adresse et y effectue un saut

Énoncé en français.