Typer la dé-sérialisation sans sérialiser les types

Grégoire Henry[1], Michel Mauny[2] et Emmanuel Chailloux[3]

1: Laboratoire Preuves, Programmation et Systèmes (PPS),
CNRS UMR 7126, Université Denis Diderot, Paris
`Gregoire.Henry@pps.jussieu.fr`
2: ENSTA, 32 Boulevard Victor, 75739 Paris cedex 15
INRIA-Rocquencourt, Domaine de Voluceau, 78153 Le Chesnay Cedex
`Michel.Mauny@{ensta,inria}.fr`
3: Laboratoire Preuves, Programmation et Systèmes (PPS),
CNRS UMR 7126, Université Pierre et Marie Curie, Paris
`Emmanuel.Chailloux@pps.jussieu.fr`

1. Introduction

La sérialisation[1] d'une valeur consiste à la représenter sous la forme d'une suite d'octets de sorte à pouvoir la sauvegarder dans un fichier pour relecture ultérieure ou la communiquer à d'autres programmes. Les langages de programmation statiquement typés cherchent à fournir, lors de la manipulation de ces valeurs dé-sérialisées, les mêmes garanties de sûreté que pour les autres valeurs. C'est pourquoi on adjoint généralement aux valeurs sérialisées une information supplémentaire permettant d'attribuer un type correct à ces valeurs lors de leur dé-sérialisation (voir par exemple [11, 17, 22, 8, 15]).

Le désavantage de l'adjonction de cette information de type est qu'elle peut rendre difficile la transmission de valeurs entre programmes écrits dans des langages différents, ou bien entre différentes versions d'un même programme. Le choix effectué par Objective Caml [16] de ne pas intégrer d'informations de type autres que celles nécessaires à la reconstruction de la valeur en mémoire dans les valeurs sérialisées est un facteur de simplicité, d'efficacité et de compacité. En conséquence, les fonctions de dé-sérialisation d'OCaml ne donnent lieu à aucune vérification, et la sécurité du typage repose sur la qualité des informations de type associées par le programmeur aux valeurs dé-sérialisées. Le programme utilisant ces fonctions de dé-sérialisation court donc le risque de rencontrer une erreur de type à l'exécution.

Dans cet article, nous remédions à ce problème en proposant une façon de donner un type statique aux fonctions de dé-sérialisation en OCaml, d'utiliser ces types statiques pour effectuer des vérifications dynamiques lors de la dé-sérialisation et nous prouvons la sûreté de ces mécanismes. À l'évidence, lorsque les valeurs considérées sont ou bien des données atomiques ou alors des données structurées (non fonctionnelles) sans partage ni cycle, la vérification de l'appartenance d'une telle valeur à un type donné est très facile à mettre en œuvre : un simple parcours récursif est suffisant. Par contre, le problème devient plus complexe lorsqu'on est en présence de partage ou de cycles, sources potentielles de polymorphisme.

Après une description des types que nous attribuons aux fonctions de dé-sérialisation à la section 2, nous décrivons informellement l'ensemble du processus de vérification à la section 3. La section 4 décrit

[1] En Anglais, on dit *serialization* ou *marshalling*, pour une mise en rang comme le ferait un officier (*marshall* ou maréchal), ou encore *pickling*, qui relève de la mise en conserve, comme en témoignent les *pickles*, petits cornichons conservés dans du vinaigre. *Marshalling* et *marshall*, s'écrivent indifféremment avec un ou deux ℓ [19].

quels sont les programmes et les valeurs que nous considérons, ainsi que la traduction des programmes en valeurs. La section 5 décrit l'algorithme de vérification, et la section 6 en énonce les propriétés, et indique quelles sont les techniques utilisées pour les prouver. La section 7 décrit brièvement les implantations prototypes qui ont été réalisées, et les sections 8 et 9 mentionnent des travaux connexes et discutent quelques axes de travail futurs.

2. Représentation dynamique des types statiques

Vérifier dynamiquement le type d'une valeur dé-sérialisée impose d'avoir accès à une représentation de ce type au moment de la dé-sérialisation. Si c'était le cas, les fonctions de dé-sérialisation de valeurs depuis une chaîne de caractères ou un fichier pourraient avoir les types[2] suivants :

$$\begin{array}{ll}\texttt{Marshal.fromString :} & \forall\alpha.\texttt{TRepr}(\alpha) \rightarrow \texttt{String} \rightarrow \alpha \\ \texttt{Marshal.fromFile :} & \forall\alpha.\texttt{TRepr}(\alpha) \rightarrow \texttt{Filename} \rightarrow \alpha\end{array}$$

où `TRepr` est un type paramétré tel que la seule valeur du type $\texttt{TRepr}(\tau)$ est une description de la représentation des valeurs du type τ. Pour τ donné, nous noterons «τ» cette valeur de type $\texttt{TRepr}(\tau)$ dans les exemples de programmes considérés dans la suite de cet article. Le programme suivant illustre l'usage d'une fonction de dé-sérialisation typée de cette façon :

```
let xs = Marshal.fromFile «List(Int)» "/tmp/f.dat"
in List.foldLeft (+) 0 xs
```

Afin d'éviter toute interaction entre les variables de types pouvant apparaître dans «τ» et dans le reste du programme, nous considérons qu'implicitement cette notation quantifie universellement les variables apparaissant dans τ et représente ainsi un schéma de type clos.

Cette possibilité de disposer des représentations de types à l'exécution sous forme de types singletons a été proposée par Crary, Weirich et Morrisett dans [6] et utilisée comme nous le faisons ici par Hicks, Weirich et Crary dans [12].

3. Présentation informelle

Avant de procéder à la présentation détaillée des langages et systèmes de types nécessaires à la description de l'algorithme et à l'énoncé de ses propriétés, nous donnons dans cette section une description informelle du processus de vérification des valeurs dé-sérialisées.

Parcourir type et valeur en parallèle Les fonctions de dé-sérialisation, munies d'une représentation de schéma de type, ont à vérifier que la représentation mémoire de la valeur v qui leur est présentée est compatible avec ce schéma de type. Cette vérification, qui utilise le type comme une contrainte que la valeur doit satisfaire, est réalisée par un parcours complet de la valeur avec, en parallèle, une expansion progressive du corps du schéma de type de sorte à exhiber le corps des définitions de constructeurs de type. Cette expansion permet, par exemple, de savoir que les seules valeurs possibles du type $\texttt{List}(\tau)$ sont représentées ou bien par un entier[3] (pour le constructeur `[]`) ou alors par un bloc marqué à 2 champs contenant respectivement des valeurs de types τ et $\texttt{List}(\tau)$. Notons que la représentation mémoire d'une valeur donnée peut être compatible avec plusieurs types non reliés : par exemple, l'entier 0 peut être indifféremment vu comme booléen, entier, ou comme le premier constructeur constant d'un type arbitraire.

[2]Nous suivons dans cet article la convention de notation préfixe des constructeurs de types où les noms de ces constructeurs commencent par une majuscule. Par exemple, le type OCaml `int list` est écrit ici `List(Int)`.

[3]Le numéro d'ordre de constructeur constant dans la liste des constructeurs constants du type.

Isomorphismes de types et polymorphisme Si, durant ce parcours, nous sommes amenés à vérifier que la valeur est de type $\forall\alpha.(\texttt{List}(\alpha)\times\texttt{List}(\alpha))$ et que la valeur courante est bien la représentation d'une paire $(v_1,\ v_2)$, nous nous ramènerons à vérifier que v_1 et v_2 sont toutes deux de type $\forall\alpha.\texttt{List}(\alpha)$, c'est-à-dire $\texttt{List}(\forall\alpha.\alpha)$. Intuitivement, une valeur de type $\forall\alpha.(\texttt{List}(\alpha)\times\texttt{List}(\alpha))$ est nécessairement une paire de listes vides, et est donc aussi de type $(\texttt{List}(\forall\alpha.\alpha)\times\texttt{List}(\forall\alpha.\alpha))$. Il s'agit là d'isomorphismes de types [5] qui nous indiquent que $\forall\alpha.(\texttt{List}(\alpha)\times\texttt{List}(\alpha))$ est équivalent à $\forall\alpha\beta.(\texttt{List}(\alpha)\times\texttt{List}(\beta))$, qui lui-même est équivalent à $(\texttt{List}(\forall\alpha.\alpha)\times\texttt{List}(\forall\beta.\beta))$. Le polymorphisme des valeurs de ce langage signifie une absence de composants (d'éléments, pour une liste, par exemple). Cette remarque permet donc de renommer les variables de type de sorte que chaque occurrence de variable soit distincte de toutes les autres, et invite à rapprocher chaque quantificateur de la variable qu'il quantifie. On transforme donc naturellement le type $\forall\alpha_1,\ldots,\alpha_n.\tau$ en $\tau[(\forall\beta.\beta)/\alpha_i]_{i=1..n}$. L'implication essentielle de cette remarque et de cette transformation sur l'algorithme de vérification est que si on est amené à vérifier qu'une (partie de) valeur est de type $\forall\alpha.\alpha$, alors l'algorithme termine en échec[4].

Au lieu d'effectuer cette transformation, nous introduisons une nouvelle constante notée $\top$, qui représentera $\forall\alpha.\alpha$, et on traduira le schéma $\forall\alpha_1,\ldots,\alpha_n.\tau$ en le type $\tau[\top/\alpha_i]_{i=1..n}$, obtenu en remplaçant toutes les variables de type apparaissant dans τ par $\top$. Nous noterons $\bar{\sigma}$ les types obtenus de cette façon. L'algorithme travaillera donc avec un terme sans variables au lieu d'un schéma de type, et devra échouer lorsqu'il aura à vérifier que la valeur courante est du type $\top$. L'équivalence entre ces deux notions de types sera formalisée par le théorème 1.

Partage Si la valeur à vérifier est un arbre (sans partage, donc), alors la vérification consiste en un simple parcours de la valeur en profondeur d'abord. Si, par contre, certains blocs sont partagés, il est tout-à-fait possible que l'un des pointeurs vers ce bloc contraigne ce dernier à être de type $\bar{\sigma}_1$ et qu'un autre pointeur lui impose d'être de type $\bar{\sigma}_2$. Si $\bar{\sigma}_1$ et $\bar{\sigma}_2$ sont incompatibles, une façon pour ce bloc d'être à la fois de type $\bar{\sigma}_1$ et $\bar{\sigma}_2$ est d'être d'un type que nous noterons $\bar{\sigma}_1 \wedge \bar{\sigma}_2$ et qui est un anti-unificateur [21, 13] de $\bar{\sigma}_1$ et $\bar{\sigma}_2$: intuitivement, pour être à la fois de type $\bar{\sigma}_1$ et $\bar{\sigma}_2$, la valeur ne doit pas avoir de composante qui soit particulière à un seul de $\bar{\sigma}_1$ ou $\bar{\sigma}_2$. Notons ici qu'une fois cet anti-unificateur calculé, nous pouvons ne parcourir qu'une seule fois ce bloc partagé et les valeurs qui en sont issues.

Cycles En présence d'un cycle, nous devons distinguer les pointeurs en provenance de l'«intérieur du cycle», c'est-à-dire de la composante fortement connexe (CFC) à laquelle appartient ce cycle, des pointeurs en provenance de l'«extérieur» de la CFC, c'est-à-dire du *contexte* de cette CFC. En effet, c'est le contexte qui fixera le type demandé à la CFC : les pointeurs en provenance du contexte sont porteurs de types qui sont anti-unifiés, comme précédemment. Une fois connues les contraintes de type imposées à chacun des nœuds racines de la CFC par son contexte, on parcourt récursivement cette CFC à partir d'une de ces racines, muni de la contrainte de type associée. Durant ce parcours, lorsqu'on rencontre des pointeurs vers les nœuds racines de la CFC, on vérifie que la contrainte associée à chacun de ces pointeurs est une instance de la contrainte émise par le contexte qui pèse sur ce nœud.

Pour résumer cette présentation informelle de l'algorithme, le parcours de la valeur va être effectué en profondeur d'abord, retardant la vérification des nœuds partagés. Pour une valeur v dont le nœud racine est partagé n fois, une fois atteints les n pointeurs qui le désignent, on parcourt la valeur v récursivement. Si on a parcouru entièrement le contexte de v en ne rencontrant que $m < n$ pointeurs vers v, alors le nœud racine de v fait partie d'un cycle : l'algorithme calcule alors ses CFC et les traite récursivement, l'un après l'autre, dans l'ordre topologique de dépendance.

[4]La vacuité du type $\forall\alpha.\alpha$ est motivée par le fait que nous ne considérons que des données structurées complètement évaluées et non pas des calculs, comme on pourrait le faire dans un langage paresseux.

4. Programmes et valeurs

Avant de décrire en détail la vérification des valeurs dé-sérialisées, fixons plus précisément les limites de cette étude : nous ne considérons dans cet article que les valeurs constituées de types de base et de types algébriques (enregistrements et unions discriminées). En particulier, on ne considèrera pas ici les objets et valeurs fonctionnelles, les laissant comme axes de travail futur : ils posent en effet des problèmes *a priori* plus difficiles que les valeurs que nous étudions. Nous ne considérons pas non plus pour le moment les types de données mutables : nous y reviendrons à la fin de cet article.

L'objectif principal est bien sûr de prouver que l'algorithme de vérification est correct (lemme 4 et théorème 2), c'est-à-dire que si la vérification de $v : \bar{\sigma}$ réussit, alors un programme e dont la sérialisation serait v possède effectivement tous les types représentés par $\bar{\sigma}$. Cela implique, par le théorème de correction du typage des programmes, que cette valeur v peut être utilisée dans un contexte d'exécution attendant une valeur de ce type.

Un autre objectif est la complétude de l'algorithme (lemme 5 et théorème 3), c'est-à-dire que tout programme e de type τ peut être sérialisé en une valeur v dont la vérification par l'algorithme réussira. Malheureusement, cette propriété n'est satisfaite que si la preuve de typage de $e : \tau$ n'utilise pas la règle de typage polymorphe de la récursion [20, 14]. Dans la pratique, l'effet de cette limitation restera probablement très marginal.

4.1. Programmes

Ce langage de valeurs correspond à un sous-ensemble des programmes source possibles. Puisque nous nous situons dans le cadre des langages à la ML, le langage source considéré, que nous appelons Val-ML, est donné par la grammaire suivante :

$$
\begin{array}{rlll}
e & ::= n & & \textit{entier} \\
 & \mid p & & \textit{variable indiquant un partage} \\
 & \mid r & & \textit{variable indiquant un cycle} \\
 & \mid (e, \ldots, e) & & \textit{n-uplet} \\
 & \mid \mathsf{C}_i & & \textit{constructeur constant} \\
 & \mid \mathsf{F}_i(e) & & \textit{constructeur non constant} \\
 & \mid \textbf{let } (p_1, \ldots, p_n) = e' \textbf{ in } e & \text{avec } p_i \in \mathit{fv}(e), \forall i = 1..n & \textit{partage} \\
 & \mid \textbf{fix } (r_1, \ldots, r_n) = e & \text{avec } r_i \in \mathit{fv}(e), \forall i = 1..n & \textit{cycle}
\end{array}
$$

où $\mathit{fv}(e)$ désigne l'ensemble des variables libres d'un programme e.

Notons que l'on impose aux déclarations **let** et **fix** des programmes d'être *utiles*, c'est-à-dire de n'introduire que des noms de variables qui sont effectivement utilisés dans leur portée. La raison en est que ces programmes seront traduits en des valeurs qui doivent être des graphes connexes. Une déclaration de variable inutilisée contribuerait à produire un graphe non connexe.

De plus, on prend soin de distinguer les pointeurs «internes» à une CFC notés r et introduits par la construction récursive **fix**, des pointeurs «externes» symbolisant simplement le partage, notés p, et introduits par la construction **let**.

Types des programmes Les expressions de type τ sont formées à partir de types de base, de variables de types, et de types paramétrés. La notion de schéma de type (notés σ) est classique, ainsi que les définitions de type, notées δ.

$$
\begin{array}{rl}
\tau & ::= \alpha \mid \mathsf{Int} \mid (\tau \times \ldots \times \tau) \ \mid T(\overrightarrow{\tau}) \\
\sigma & ::= \forall \alpha . \tau \mid \tau \\
\delta & ::= T(\overrightarrow{\alpha}) = [\ \mathsf{C}_1 \mid \ldots \mid \mathsf{C}_n \mid \mathsf{F}_1(\tau) \mid \ldots \mid \mathsf{F}_k(\tau)\ ]
\end{array}
$$

Les types de n-uplets sont notés comme des produits cartésiens d'arité variable. Pour simplifier la présentation, ces définitions de types sont considérées comme mutuellement récursives, et les constructeurs de données qu'elles introduisent sont supposés distincts les uns des autres, afin qu'un constructeur de données identifie de façon non-ambiguë un constructeur de type. Ici, $\overrightarrow{\alpha}$ représente les paramètres formels du constructeur de types T, les C_i sont ses constructeurs constants et les $\mathsf{F}_i(\_)$ ses constructeurs non constants[5].

Pour un programme donné, l'ensemble de ses définitions de types est noté Δ. En notant θ la substitution qui associe les $\overrightarrow{\tau}$ aux $\overrightarrow{\alpha}$, on pose, pour $T(\overrightarrow{\alpha})$ défini dans Δ :

$$\Delta(T(\overrightarrow{\tau})) = [\ \mathsf{C}_1 \mid \ldots \mid \mathsf{C}_n \mid \mathsf{F}_1(\theta(\tau_1)) \mid \ldots \mid \mathsf{F}_k(\theta(\tau_k))\ ]$$

ce qui nous permet d'accéder directement à la définition instanciée de T.

Environnement de typage Les environnements de typage sont notés Γ et définis par :

$$\Gamma \ ::= \ \emptyset \mid \Gamma \oplus \{p : \sigma\} \mid \Gamma \oplus \{r : \sigma\}$$

On utilisera souvent Γ comme une fonction associant à un nom de pointeur q (où q est un pointeur p ou r) un schéma de type $\Gamma(q)$. On notera $dom(\Gamma)$ l'ensemble des variables q pour lesquels $\Gamma(q)$ est défini et, $\Gamma \backslash \{q_1, \ldots, q_n\}$ la restriction de Γ à $dom(\Gamma) \backslash \{q_1, \ldots, q_n\}$. Enfin, on abrégera la généralisation d'un type τ dans un environnement Γ en notant :

$$gen(\Gamma, \tau) \ = \ \forall \overrightarrow{fv(\tau) \backslash fv(\Gamma)} \cdot \tau$$

Instanciation (polymorphe) On notera :
- $\tau \leq \sigma$ (τ est une instance polymorphe de σ), si σ s'écrit $\forall \overrightarrow{\alpha}.\tau'$ et s'il existe une substitution θ telle que $dom(\theta) = \{\overrightarrow{\alpha}\}$ et $\tau = \theta(\tau')$;
- $\sigma \preceq \sigma'$ si toute instance polymorphe de σ est aussi instance polymorphe de σ' (en d'autres termes, σ' contient plus d'instances de type que σ).

Étendant la relation d'instanciation polymorphe aux environnements de typage, on écrira $\Gamma \preceq \Gamma'$ si :
- $dom(\Gamma) = dom(\Gamma')$;
- $\forall p \in dom(\Gamma),\ \Gamma(p) \preceq \Gamma'(p)$;
- $\forall r \in dom(\Gamma),\ \Gamma(r) =_\alpha \Gamma'(r)$.

Remarquons que les pointeurs récursifs reçoivent un traitement spécifique : pour instancier un environnement, on ne peut instancier que les types associés à des pointeurs non récursifs (notés p). Les types associés à des pointeurs récursifs (notés r) doivent, quant à eux, rester inchangés à un renommage près, comme de coutume. Cette distinction est due au traitement différent reçu par les pointeurs de partage — pour lesquels des types distincts sont anti-unifiés — de celui reçu par les pointeurs récursifs, dont on vérifie simplement que le type est une instance du type du cycle dont ils sont racine.

4.2. Valeurs

Le langage des valeurs que nous considérons peut être représenté par la grammaire suivante. Les blocs représentent de la mémoire allouée et sont dotés d'une marque entière i, d'une taille n et de n valeurs contenues dans chacun des champs. Tous les pointeurs sont déclarés : ils indiquent la présence de partage ou de cycles effectifs. Aucun pointeur déclaré n'est superflu. La construction de ces valeurs à partir de leur représentation sérialisée est décrite à la section 7.

$$\begin{array}{lll}
v \ ::= & n \mid p \mid r & \textit{entier, pointeurs et pointeurs récursifs} \\
& \mid \mathsf{Block}(i, v_1, \ldots, v_n) & \textit{bloc alloué de marque } i \geq 1 \textit{ et d'arité } n \geq 1 \\
& \mid \mathsf{let}\ (p_1, \ldots, p_n) = v\ \mathsf{in}\ w & \textit{partage} \\
& \mid \mathsf{fix}\ (r_1, \ldots, r_n) = v & \textit{cycles (composantes fortement connexes)}
\end{array}$$

[5]Nous ne considérons ici que les constructeurs de données constants ou unaires afin de simplifier la présentation. Les constructeurs n-aires induisent en effet des cas de preuve partiellement redondants avec le cas des n-uplets.

Types des valeurs Les types des valeurs sont construits comme les types des programmes où une constante notée $\top$ représentant le type vide, aura pris la place des variables de types. On les notera comme des termes de la grammaire suivante :

$$\bar{\sigma} \ ::= \ \top \mid \mathsf{Int} \mid (\bar{\sigma}_1 \times \ldots \times \bar{\sigma}_n) \mid T(\bar{\sigma}_1, \ldots, \bar{\sigma}_n)$$

On les définit formellement comme les représentants de classes d'équivalence de schémas de types clos. La relation d'équivalence est notée $\sim$ et est définie par :

$$\forall \alpha_1, \ldots, \alpha_n.\tau \sim \forall \beta_1, \ldots, \beta_m.\tau' \text{ si et seulement si } \tau[\gamma/\alpha_1, \ldots, \alpha_n] = \tau'[\gamma/\beta_1, \ldots, \beta_m]$$

où γ est une variable «fraîche». En d'autres termes, deux schémas sont équivalents s'ils ne diffèrent qu'aux occurrences de variables. Il est immédiat que cette relation est une relation d'équivalence.

On se convainc aisément que chacune de ces classes d'équivalence contient des schémas de types qui sont isomorphes au sens que nous avons donné à la section 3, et le théorème 1 démontre que, pour une expression close e typable par un schéma de type clos σ, les différents membres de la classe de σ sont aussi des types valides pour e.

Le représentant de la classe d'un schéma σ peut être calculé par la fonction *univ* ainsi définie :

$$\begin{aligned}
univ(\forall \alpha.\sigma) &= univ(\sigma)\\
univ(\alpha) &= \top\\
univ(\mathsf{Int}) &= \mathsf{Int}\\
univ((\tau_1 \times \ldots \times \tau_n)) &= (univ(\tau_1) \times \ldots \times univ(\tau_n))\\
univ(T(\tau_1, \ldots, \tau_n)) &= T(univ(\tau_1), \ldots, univ(\tau_n))
\end{aligned}$$

Inversement, on notera $schema(\bar{\sigma})$ le plus général des schémas de type de la classe représenté par $\bar{\sigma}$.

On écrira $\bar{\sigma} \preceq \bar{\sigma}'$ si $schema(\bar{\sigma}) \preceq schema(\bar{\sigma}')$. Et en notant $\bar{\Gamma}$, $\bar{\Gamma}'$, ..., les environnements de typage composés d'assertions de la forme $q : \bar{\sigma}$, on étend cette relation d'instanciation polymorphe aux environnements de la même manière que pour les classes de types : en exigeant l'égalité sur les pointeurs récursifs.

4.3. Traduction des programmes en valeurs

Le passage d'un programme à une valeur est assuré par la fonction de traduction suivante :

$$\begin{aligned}
[\![\, n \,]\!] &= n\\
[\![\, p \,]\!] &= p\\
[\![\, r \,]\!] &= r\\
[\![\, (e_1, \ldots, e_n) \,]\!] &= \mathsf{Block}(0, [\![\, e_1 \,]\!], \ldots, [\![\, e_n \,]\!])\\
[\![\, \mathsf{C}_i \,]\!] &= i\\
[\![\, \mathsf{F}_i(e) \,]\!] &= \mathsf{Block}(i, [\![\, e \,]\!])\\
[\![\, \textbf{let } (p_1, \ldots, p_n) = e' \textbf{ in } e \,]\!] &= \mathsf{let}\ (p_1, \ldots, p_n) = [\![\, e' \,]\!] \ \mathsf{in}\ [\![\, e \,]\!]\\
[\![\, \textbf{fix } (r_1, \ldots, r_n) = e \,]\!] &= \mathsf{fix}\ (r_1, \ldots, r_n) = [\![\, e \,]\!]
\end{aligned}$$

Cette traduction identifie les entiers et les constructeurs constants de même rang, ainsi que les constructeurs fonctionnels de même rang et de même arité. La vérification de types à laquelle on procède ici a pour but de lever les ambiguïtés introduites par cette traduction.

4.4. Anti-unification sans mémoire

L'anti-unification classique de deux termes du premier ordre produit un troisième terme, appelé anti-unificateur, dont les deux premiers sont des instances. Parmi les anti-unificateurs de deux termes donnés, le meilleur est celui qui est le moins général, instance de tous les autres. Classiquement, l'anti-unification produit donc des termes avec variables. Nous en donnons ici une définition légèrement différente, que nous appelons *unification sans mémoire* et qui n'utilise pas de variables mais la constante $\top$, produite lorsque deux termes à anti-unifier entrent en conflit. Là où l'anti-unification classique produirait plusieurs fois la même variable, mémorisant ainsi plusieurs occurrences d'un même conflit, nous produisons plusieurs occurrences du terme $\top$ qui pourront être instanciées indépendamment les unes des autres, conformément à l'intuition que nous procure les isomorphismes de types mentionnés à la section 3. L'anti-unification sans mémoire de $\bar{\sigma}_1$ et $\bar{\sigma}_2$ est notée $\bar{\sigma}_1 \wedge \bar{\sigma}_2$, et est définie par :

$$\begin{array}{rcll} \mathsf{Int} \wedge \mathsf{Int} & = & \mathsf{Int} & \\ (\bar{\sigma}_1 \times \ldots \times \bar{\sigma}_n) \wedge (\bar{\sigma}'_1 \times \ldots \times \bar{\sigma}'_n) & = & (\bar{\sigma}_1 \wedge \bar{\sigma}'_1 \times \ldots \times \bar{\sigma}_n \wedge \bar{\sigma}'_n) & \\ T(\bar{\sigma}_1, \ldots, \bar{\sigma}_n) \wedge T'(\bar{\sigma}'_1, \ldots, \bar{\sigma}'_n) & = & T(\bar{\sigma}_1 \wedge \bar{\sigma}'_1, \ldots, \bar{\sigma}_n \wedge \bar{\sigma}'_n) & \textit{si } T = T' \\ \bar{\sigma}_1 \wedge \bar{\sigma}_2 & = & \top & \textit{dans tous les autres cas} \end{array}$$

Dans la suite, ne faisant plus référence à l'anti-unification classique, nous écrirons simplement *anti-unification* pour dénoter l'anti-unification sans mémoire.

On ajoute à notre algèbre de types un élément noté $\perp$, neutre pour l'anti-unification. Cet élément sera utilisé comme hypothèse de type initiale pour les pointeurs lorsque l'algorithme entre dans leur portée, et sera nécessairement anti-unifié avec le type d'au moins une occurrence de ce pointeur : en effet, les contraintes de formation des programmes, qui imposent aux variables liées d'apparaître libres au noins une fois dans leur portée (aucune déclaration n'est inutile) sont, par traduction, transférées aux valeurs. Il en résulte que $\perp$ ne sera jamais utilisé comme argument de la fonction de vérification.

5. *Check*, algorithme de vérification de types

La vérification est représentée par le calcul de $\mathit{Check}(\bar{\Gamma}, \bar{\sigma}, v)$ où $\bar{\Gamma}$ est un environnement, et $\bar{\sigma}$ et v sont respectivement le type et la valeur à vérifier. Un appel $\mathit{Check}(\bar{\Gamma}, \bar{\sigma}, v)$ produira un couple $(e, \bar{\Gamma}')$ ou bien échouera. En cas de succès, $\bar{\Gamma}'$ est une généralisation de $\bar{\Gamma}$ (ce que montrera le lemme 3), et e est un programme dont la représentation est la valeur v. Un appel initial à *Check* est de la forme $\mathit{Check}(\emptyset, \bar{\sigma}, v)$ et produira une paire $(e, \emptyset)$ en cas de succès. Un invariant de cette fonction est que $\mathit{fv}(v) \subseteq \mathit{dom}(\bar{\Gamma})$: en d'autres termes tous les pointeurs libres de v se voient attribuer un type (éventuellement $\perp$) dans $\bar{\Gamma}$. Il va de soi que dans une implémentation effective, cet algorithme se bornera à parcourir la valeur examinée et la rendre en résultat si la vérification réussit.

On notera $\bar{\Gamma} \otimes \{p : \bar{\sigma}\}$, l'environnement $\bar{\Gamma}'$ tel que $\bar{\Gamma}'(p) = \bar{\Gamma}(p) \wedge \bar{\sigma}$ et $\bar{\Gamma}'(p') = \bar{\Gamma}(p')$ pour $p' \neq p$.

$\mathit{Check}(\bar{\Gamma}, \mathsf{Int}, i) = (i, \bar{\Gamma})$

$\mathit{Check}(\bar{\Gamma}, T(\bar{\sigma}_1, \ldots, \bar{\sigma}_m), i) =$
  let $[\ \mathsf{C}_1 \mid \ldots \mid \mathsf{C}_n \mid \mathsf{F}_1(\bar{\sigma}'_1) \mid \ldots \mid \mathsf{F}_k(\bar{\sigma}'_k)\ ] = \Delta(T(\bar{\sigma}_1, \ldots, \bar{\sigma}_m))$ **in**
  if $0 < i \leq n$ **then** $(\mathsf{C}_i, \bar{\Gamma})$ **else Failure**

$\mathit{Check}(\bar{\Gamma}, T(\bar{\sigma}_1, \ldots, \bar{\sigma}_m), \mathsf{Block}(i, v')) =$
  let $[\ \mathsf{C}_1 \mid \ldots \mid \mathsf{C}_n \mid \mathsf{F}_1(\bar{\sigma}'_1) \mid \ldots \mid \mathsf{F}_k(\bar{\sigma}'_k)\ ] = \Delta(T(\bar{\sigma}_1, \ldots, \bar{\sigma}_m))$ **in**
  if $0 < i \leq k$ **then**
    let $(e', \bar{\Gamma}') = \mathit{Check}(\bar{\Gamma}, \bar{\sigma}'_i, v')$ **in**
    $(\mathsf{F}_i(e'), \bar{\Gamma}')$
  else
    Failure

$\mathit{Check}(\bar{\Gamma}, (\bar{\sigma}_1 \times \ldots \times \bar{\sigma}_n), \mathsf{Block}(0, v_1, \ldots, v_n)) =$
  let $(e_1, \bar{\Gamma}_1) = \mathit{Check}(\bar{\Gamma}, \bar{\sigma}_1, v_1)$ **in**
  $\ldots$
  let $(e_n, \bar{\Gamma}_n) = \mathit{Check}(\bar{\Gamma}_{n-1}, \bar{\sigma}_n, v_n)$ **in**
  $((e_1, \ldots, e_n), \bar{\Gamma}_n)$

$\mathit{Check}(\bar{\Gamma}, \bar{\sigma}, p) =$
    $(p, (\bar{\Gamma} \otimes \{p : \bar{\sigma}\}))$

$\mathit{Check}(\bar{\Gamma}, \bar{\sigma}, r) =$
  if $\bar{\sigma} \preceq \bar{\Gamma}(r)$ **then** $(r, \bar{\Gamma})$ **else Failure**

$\mathit{Check}(\bar{\Gamma}, \bar{\sigma}, (\mathsf{let}\ (p_1, \ldots, p_n) = v\ \mathsf{in}\ w)) =$
  let $(e, \bar{\Gamma}') = \mathit{Check}(\bar{\Gamma} \oplus \{p_i : \bot\}_{i=1..n}, \bar{\sigma}, w)$ **in**
  let $(e', \bar{\Gamma}'') = \mathit{Check}(\bar{\Gamma}' \backslash \{p_1, \ldots, p_n\}, (\bar{\Gamma}'(p_1), \ldots, \bar{\Gamma}'(p_n)), v)$ **in**
  $((\mathbf{let}\ (p_1, \ldots, p_n) = e'\ \mathbf{in}\ e), \bar{\Gamma}'')$

$\mathit{Check}(\bar{\Gamma}, (\bar{\sigma}_1, \ldots, \bar{\sigma}_n), (\mathsf{fix}\ (r_1, \ldots, r_n) = v)) =$
  let $\bar{\Gamma}' = \bar{\Gamma} \oplus \{r_i : \bar{\sigma}_i\}_{i=1..n}$ **in**
  let $(e, \bar{\Gamma}'') = \mathit{Check}(\bar{\Gamma}', (\bar{\sigma}_1, \ldots, \bar{\sigma}_n), v)$ **in**
  $((\mathbf{fix}\ (r_1, \ldots, r_n) = e), \bar{\Gamma}'' \setminus \{r_i\}_{i=1..n})$

Dans tous les autres cas :
  $\mathit{Check}(\bar{\Gamma}, \bar{\sigma}, v) = \mathbf{Failure}$

6. Propriétés de l'algorithme

Nous énonçons dans cette section les propriétés que satisfait l'algorithme de vérification. Nous en donnons aussi les schémas de preuves, mais invitons le lecteur à se reporter à [10], où il pourra trouver les preuves complètes.

6.1. Typage des programmes

La discipline de typage à laquelle sont soumis les programmes est la discipline habituelle de typage de ML avec récursion polymorphe [20]. Puisque nous ne considérons pas les valeurs mutables, la généralisation n'est pas soumise à restriction.

6.1.1. $\vdash^{ML^{rec}}$, typage classique des programmes, avec récursion polymorphe

$$\Gamma \vdash^{ML^{rec}} n : \mathsf{Int} \qquad \frac{\tau \leq \Gamma(p)}{\Gamma \vdash^{ML^{rec}} p : \tau} \qquad \frac{\tau \leq \Gamma(r)}{\Gamma \vdash^{ML^{rec}} r : \tau} \qquad \frac{\forall i = 1..n,\ \Gamma \vdash^{ML^{rec}} e_i : \tau_i}{\Gamma \vdash^{ML^{rec}} (e_1, \ldots, e_n) : (\tau_1 \times \ldots \times \tau_n)}$$

$$\frac{\mathsf{C}_i \in \Delta(T(\overrightarrow{\tau}))}{\Gamma \vdash^{ML^{rec}} \mathsf{C}_i : T(\overrightarrow{\tau})} \qquad \frac{\mathsf{F}_i(\tau') \in \Delta(T(\overrightarrow{\tau})) \qquad \Gamma \vdash^{ML^{rec}} e : \tau'}{\Gamma \vdash^{ML^{rec}} \mathsf{F}_i(e) : T(\overrightarrow{\tau})}$$

$$\frac{\Gamma \vdash^{ML^{rec}} e' : (\tau_1 \times \ldots \times \tau_n) \qquad \Gamma \oplus \{p_i : gen(\Gamma, \tau_i)\}_{i=1..n} \vdash^{ML^{rec}} e : \tau}{\Gamma \vdash^{ML^{rec}} \mathbf{let}\ (p_1, \ldots, p_n) = e'\ \mathbf{in}\ e : \tau}$$

$$\frac{\Gamma \oplus \{r_i : gen(\Gamma, \tau_i')\}_{i=1..n} \vdash^{ML^{rec}} e : (\tau_1' \times \ldots \times \tau_n')}{\Gamma \vdash^{ML^{rec}} \mathbf{fix}\ (r_1, \ldots, r_n) = e : (\tau_1 \times \ldots \times \tau_n)} \quad (\forall i = 1..n,\ \tau_i \leq gen(\Gamma, \tau_i'))$$

Rappelons [14] que si $\Gamma' \succeq \Gamma$ et $\Gamma \vdash^{ML^{rec}} e : \tau$ alors $\Gamma' \vdash^{ML^{rec}} e : \tau$.

6.1.2. $\vdash^{\top}$, un système de types alternatif

Nous introduisons ici une manière de typer nos programmes *modulo* isomorphismes de types, associant à un programme e un type $\bar{\sigma}$, où les variables de types ont été remplacées par la constante

$\top$.

On définit le prédicat $\bar{\Gamma} \vdash^{\top} e : \bar{\sigma}$ par le système de règles suivant :

$$\frac{}{\bar{\Gamma} \vdash^{\top} n : \mathsf{Int}} \qquad \frac{\bar{\sigma} \preceq \bar{\Gamma}(p)}{\bar{\Gamma} \vdash^{\top} p : \bar{\sigma}} \qquad \frac{\bar{\sigma} \preceq \bar{\Gamma}(r)}{\bar{\Gamma} \vdash^{\top} r : \bar{\sigma}} \qquad \frac{\forall i = 1..n,\ \bar{\Gamma} \vdash^{\top} e_i : \bar{\sigma}_i}{\bar{\Gamma} \vdash^{\top} (e_1, \ldots, e_n) : (\bar{\sigma}_1 \times \ldots \times \tau_n)} \qquad \frac{\mathsf{C}_i \in \Delta(T(\overrightarrow{\sigma}))}{\bar{\Gamma} \vdash^{\top} \mathsf{C}_i : T(\overrightarrow{\sigma})}$$

$$\frac{\mathsf{F}_i(\bar{\sigma}') \in \Delta(T(\overrightarrow{\sigma})) \qquad \bar{\Gamma} \vdash^{\top} e : \bar{\sigma}'}{\bar{\Gamma} \vdash^{\top} \mathsf{F}_i(e) : T(\overrightarrow{\sigma})} \qquad \frac{\bar{\Gamma} \vdash^{\top} e' : (\bar{\sigma}_1 \times \ldots \times \bar{\sigma}_n) \qquad \bar{\Gamma} \oplus \{p_i : \bar{\sigma}_i\}_{i=1..n} \vdash^{\top} e : \bar{\sigma}}{\bar{\Gamma} \vdash^{\top} \mathbf{let}\ (p_1, \ldots, p_n) = e'\ \mathbf{in}\ e : \bar{\sigma}}$$

$$\frac{\bar{\Gamma} \oplus \{r_i : \bar{\sigma}'_i\}_{i=1..n} \vdash^{\top} e : (\bar{\sigma}'_1 \times \ldots \times \bar{\sigma}'_n)}{\bar{\Gamma} \vdash^{\top} \mathbf{fix}\ (r_1, \ldots, r_n) = e : (\bar{\sigma}_1 \times \ldots \times \bar{\sigma}_n)} \quad (\forall i = 1..n,\ \bar{\sigma}_i \preceq \bar{\sigma}'_i)$$

Lemme 1 *Si $\bar{\Gamma}' \succeq \bar{\Gamma}$ et $\bar{\Gamma} \vdash^{\top} e : \bar{\sigma}$, alors $\bar{\Gamma}' \vdash^{\top} e : \bar{\sigma}$.*

Schéma de preuve Par induction sur l'arbre de typage de $\bar{\Gamma} \vdash^{\top} e : \bar{\sigma}$.

6.2. Équivalence entre $\vdash^{ML^{rec}}$ et $\vdash^{\top}$

Lemme 2 *Soient un programme e, un type τ_e et un environnement Γ. Si $\Gamma \vdash^{ML^{rec}} e : \tau_e$ alors pour $\bar{\Gamma} = univ(\Gamma)$ et $\bar{\sigma}_e = univ(\tau_e)$ on a : $\bar{\Gamma} \vdash^{\top} e : \bar{\sigma}_e$.*

Schéma de preuve Par réécriture systématique de l'arbre de typage de $\Gamma \vdash^{ML^{rec}} e : \tau_e$.

Théorème 1 (Équivalence) *Soient un programme e, un type $\bar{\sigma}$ et un environnement $\bar{\Gamma}$. Si $\bar{\Gamma} \vdash^{\top} e : \bar{\sigma}$ alors pour $\Gamma = schema(\bar{\Gamma})$ et pour tout $\tau \leq schema(\bar{\sigma})$ on a : $\Gamma \vdash^{ML^{rec}} e : \tau$.*

Schéma de preuve Par induction sur la dérivation du jugement $\bar{\Gamma} \vdash^{\top} e : \bar{\sigma}$ et par cas selon la dernière règle utilisée.

6.3. Correction

Lemme 3 *Soient une valeur v, un type $\bar{\sigma}$, et un environnement de type $\bar{\Gamma}$ tels que $fv(v) \subseteq dom(\bar{\Gamma})$. Si $Check(\bar{\Gamma}, \bar{\sigma}, v) = (e, \bar{\Gamma}')$ alors :*

1. *$dom(\bar{\Gamma}') = dom(\bar{\Gamma})$,*
2. *$\bar{\Gamma}' \succeq \bar{\Gamma}$,*
3. *$[\![e]\!] = v$.*

Schéma de preuve Par induction sur les appels récursifs de $Check(\bar{\Gamma}, \bar{\sigma}, v)$.

Lemme 4 *Soient une valeur v, un type $\bar{\sigma}$, et un environnement de type $\bar{\Gamma}$ tels que $fv(v) \subseteq dom(\bar{\Gamma})$. Si $Check(\bar{\Gamma}, \bar{\sigma}, v) = (e, \bar{\Gamma}')$ alors $\bar{\Gamma}' \vdash^{\top} e : \bar{\sigma}$.*

Schéma de preuve Par induction sur les appels récursifs de $Check(\bar{\Gamma}, \bar{\sigma}, v)$.

Théorème 2 (Correction) *Soient une valeur v, et un type $\bar{\sigma}$. Si $Check(\emptyset, \bar{\sigma}, v)$ réussit en calculant un programme e, alors : $\emptyset \vdash^{\top} e : \bar{\sigma}$.*

Preuve Le lemme 3 nous indique $Check(\emptyset, \bar{\sigma}, v) = (e, \emptyset)$. C'est ensuite une application directe du lemme 4.

Le théorème 2 et le lemme 4 nous indiquent que si la vérification de v réussit en $\bar{\sigma}$, alors l'algorithme produit un programme e dont la valeur est v et qui est typable par $\bar{\sigma}$. Cela rend bien évidemment inutile la production de e dans une implémentation effective de l'algorithme, qui peut se borner à rendre v en résultat en cas de succès.

6.4. Complétude

Notre algorithme n'est pas complet dans le sens où toute expression admettant une dérivation de typage dans $\vdash^{ML^{rec}}$ ne sera pas forcément acceptée par notre algorithme. En effet, les définitions de types non réguliers fournissent des contre-exemples, puisque certaines définitions de structures de données récursives nécessitent un typage polymorphe de la récursion pour être acceptées en ML. Soit par exemple la définition de type :

$$\mathsf{Nest}(\alpha) = [\ \ldots \mid \mathsf{B}(\mathsf{Nest}(\alpha \times \alpha)) \mid \ldots\ ]$$

et considérons la valeur **fix** $(r) = \mathsf{B}(r)$. On montre aisément que cette valeur peut avoir le type $\mathsf{Nest}(\mathsf{Int})$:

$$\dfrac{\dfrac{\{r : \forall\alpha.\mathsf{Nest}(\alpha)\} \vdash^{ML^{rec}} r : \mathsf{Nest}(\alpha \times \alpha)}{\{r : \forall\alpha.\mathsf{Nest}(\alpha)\} \vdash^{ML^{rec}} \mathsf{B}\,(r) : \mathsf{Nest}(\alpha)}}{\emptyset \vdash^{ML^{rec}} \mathbf{fix}\ (r) = \mathsf{B}\,(r) : \mathsf{Nest}(\mathsf{Int})} \quad (\mathsf{Nest}(\mathsf{Int}) \leq \forall\alpha.\mathsf{Nest}(\alpha))$$

Cependant, lorsqu'on demande à l'algorithme de vérifier cette même assertion, celui-ci va échouer en tentant de vérifier que $\mathsf{Nest}(\mathsf{Int} \times \mathsf{Int}) \preceq \mathsf{Nest}(\mathsf{Int})$.

Nous allons montrer par contre que notre algorithme est complet pour toute valeur typable sans avoir recours à la récursion polymorphe. Soit $\vdash^{ML}$ le système obtenu de $\vdash^{ML^{rec}}$ en y remplaçant la règle de typage polymorphe de la récursion par la règle classique de typage (monomorphe) de la récursion :

$$\dfrac{\Gamma \oplus \{r_i : \tau_i\}_{i=1..n} \vdash^{ML} e : (\tau_1 \times \ldots \times \tau_n)}{\Gamma \vdash^{ML} \mathbf{fix}\ (r_1, \ldots, r_n) = e : (\tau_1 \times \ldots \times \tau_n)}$$

Lemme 5 *Soient une expression e, un type τ et un environnement Γ tels que $fv(e) \subseteq dom(\Gamma)$. Soient une substitution θ et un type $\bar{\sigma}$ tels que $dom(\theta) = fv(\Gamma) \cup fv(\tau)$, et $fv(img(\theta)) \# dom(\theta)$. Posons $\bar{\sigma} = univ(\theta(\tau))$.*

Si $\Gamma \vdash^{ML} e : \tau$ alors pour tout $\Gamma' \preceq \theta(\Gamma)$, on a $Check(univ(\Gamma'), \bar{\sigma}, [\![e]\!]) = (e, \bar{\Gamma}'')$ où $\bar{\Gamma}'' = univ(\Gamma'')$ avec $\Gamma'' \preceq \theta(\Gamma)$.

Schéma de preuve Par induction sur la dérivation de typage $\Gamma \vdash^{ML} e : \tau$, et par cas selon la dernière règle utilisée.

Théorème 3 (Complétude) *Soient une expression close e et un type τ. Si $\emptyset \vdash^{ML} e : \tau$, alors $Check(\emptyset, univ(\tau), [\![e]\!])$ réussit.*

Preuve C'est un cas particulier du lemme 5.

6.5. Complexité du parcours

Pour une valeur sans cycle, l'algorithme de parcours est linéaire en la taille N du programme la représentant. En effet, le choix d'anti-unifier le type de toutes les références à un bloc partagé avant de parcourir ce dernier permet de ne parcourir qu'une seule fois l'ensemble de la valeur. En présence de cycles il faut ajouter au coût de ce parcours celui de l'identification et du tri des CFC, qui est dans le pire des cas de l'ordre de $N \times P$ où P est la profondeur maximale d'imbrication de constructions **fix**.

7. Implantation

Une implantation prototype de cet algorithme de vérification a été réalisée, ainsi que de certains éléments nécessaires à son intégration en OCaml. Cette section décrit quelques-uns des points importants de cette mise en œuvre ; le lecteur intéressé pourra se référer à [10] pour y trouver une description plus précise.

Représentation des types Les types abstraits et la compilation séparée font que lorsqu'on compile une valeur «τ», on ne dispose pas statiquement de toutes les informations nécessaires à la reconnaissance des valeurs de type τ. On résoud ce problème en compilant chaque déclaration de type comme une fonction ayant autant de paramètres que le constructeur de types. Ainsi, dans une expression «τ», une référence à un type abstrait sera compilée comme un appel à la fonction correspondante dont le code ne sera accessible que lors de l'exécution, puisque fourni par l'implémentation de l'interface exportant ce type. Cette solution présente l'avantage de laisser inchangées les signatures de modules et de ne modifier que les implémentations.

Détection du partage et tri topologique Nous utilisons pour la recherche des CFC et leurs tris topologiques l'algorithme proposé par Tarjan [23]. Les informations sur le partage peuvent être accessibles en temps constant en encapsulant chaque valeur partagée dans un bloc spécial lors de la reconstruction en mémoire de la valeur dé-sérialisée. Cette capsule permet d'accéder aux informations nécessaires à l'algorithme de recherche de CFC, ainsi qu'au résultat de l'anti-unification des contraintes de type déjà rencontrées pour le bloc encapsulé. Chaque référence sur cette capsule n'étant suivie qu'une seule fois, le graphe mémoire réel de la valeur peut être reconstitué incrémentalement en remplaçant systématiquement la référence sur la capsule par une référence sur le bloc encapsulé lors du parcours de vérification.

L'algorithme de parcours d'une valeur, présenté à la section 3, parcourt les blocs du graphe mémoire d'une valeur dans le même ordre que l'algorithme *Check*. Chaque bloc partagé introduit une nouvelle construction **let** $(p) = e$ **in** e' où p est une variable fraîche, e' le contexte déjà exploré et e la valeur représentée par ce bloc. Chaque CFC est représentée par une construction **fix** introduisant une variable pour chacune de ses racines. Des constructions **fix** imbriquées représentent des CFC internes à une CFC.

8. Travaux connexes

Les solutions apportées jusqu'à maintenant à ce problème se résument à sérialiser non seulement une valeur, mais aussi son type sous une forme ou sous une autre. Au début des années 1980, Herlihy et Liskov ont proposé une méthode de transmission de valeurs s'appliquant aux types abstraits dans le langage CLU [18] : les valeurs sont sérialisées avec leur type et des fonctions d'encodage et de décodage spécifiques fournies par les types abstraits sont utilisées à la place des mécanismes généraux de sérialisation. Leroy et Mauny [17] ont étudié l'introduction de valeurs à types dynamiques en ML, étendant ainsi une proposition initiale de Cardelli dans le langage Amber [3]. Les valeurs à types dynamiques (ou plus simplement *dynamiques*) sont des paires composées d'une valeur v et d'un type τ telles que $v : \tau$. La création de dynamiques nécessite la collaboration du compilateur, qui inclut le type statique τ de la valeur v afin de créer le dynamique (v, τ). Pour passer du dynamique (v, τ) à la valeur typée $v : \tau$ (le «déconstruire»), il est nécessaire d'avoir recours à un mécanisme particulier de filtrage qui réalise des tests de type sur les dynamiques. Puisque les dynamiques ont tous le même type, ils sont à même d'être manipulés par des fonctions de lecture et d'écriture typées. Les fonctions de sérialisation et de dé-sérialisation ont alors respectivement les types `Dynamic` $\rightarrow$ `Filename` $\rightarrow$ `Unit` et `Filename` $\rightarrow$ `Dynamic`. On se retrouve alors dans une situation similaire au mécanisme de sérialisation actuellement implanté en Java.

Les travaux de Dubois, Rouaix et Weis sur le polymorphisme générique [7] ont permis à Furuse et Weis de concevoir une forme de valeurs à types dynamiques, et de proposer des fonctions de lecture et d'écriture de valeurs avec leurs types. L'information de type qui est écrite est une empreinte cryptographique du type d'origine de la valeur, *modulo* renommage de labels et de constructeurs, fournissant ainsi une certaine souplesse grâce à la possibilité de renommage, et une efficacité des tests de type — nécessaires à la déconstruction des dynamiques — puisque seules des empreintes doivent être comparées.

Leifer, Peskine, Sewell et Wansbrough [15] s'intéressent quant à eux à la préservation de l'abstraction, renonçant donc à la souplesse mentionnée plus haut. Pour garantir la préservation de l'abstraction, ils associent aux valeurs l'empreinte cryptographique des définitions de leur type et de leur contexte (c'est-à-dire essentiellement le texte du module les contenant ainsi que les modules qui y sont importés).

Du côté du monde objet à la Java, la dé-sérialisation d'un objet produit une instance du type `Object`. C'est ensuite le mécanisme de *downcasting* qui en assurera la spécialisation à la demande.

L'approche que nous suivons se différencie clairement de ces travaux, en ce que, d'une part, nous considérons la dé-sérialisation de valeurs qui ne portent aucune information de type, et d'autre part, la vérification de typage que nous effectuons allie garantie et souplesse.

Le problème de la reconstruction dynamique du type des valeurs dans les langages statiquement typés comme ML a surtout été étudié dans le cadre de l'interaction entre l'optimisation de la représentation des données et la gestion automatique de la mémoire (voir par exemple [2, 9, 1, 24, 4]), ainsi que dans le but de concevoir des outils de mise au point. La difficulté majeure à surmonter dans ce contexte est de retrouver les *instances* d'utilisation de valeurs polymorphes. En effet, un gestionnaire mémoire ou un *debugger* est amené à manipuler des valeurs résultant de calculs qui ont impliqué des utilisations d'instances particulières de valeurs polymorphes, et la reconstruction de ces instanciations est essentielle à l'obtention d'informations de types fiables sur la mémoire. Par exemple, une liste paramètre de la fonction `List.length` a pour type statique `List(`α`)` alors qu'un argument effectif de cette même fonction peut être de type `List(Int)` ou bien `List(Bool`$\times$`Float)`. Le problème de la reconstruction dynamique de types pour la gestion de la mémoire ou la conception d'outils de mise au point consiste donc à retrouver cette instance, en reconstituant l'histoire du calcul.

En règle générale, cette reconstruction impose au programme de stocker suffisamment d'informations de type à l'exécution pour pouvoir reconstruire cet historique, pour garder trace des applications de fonctions polymorphes et prendre en compte les mutations et les levées d'exceptions. Par contre, ces travaux font l'hypothèse que cette information de type est correcte : on reconstruit donc le type d'un programme à un moment de son exécution mais on ne procède à aucune vérification.

Le travail présenté ici ne nécessite pas de reconstruire ainsi le type d'un programme, mais vise au contraire à *vérifier* qu'une valeur appartient bien à un type.

9. Discussion et travaux futurs

Un traitement correct des données mutables est indispensable avant d'envisager toute intégration dans un langage de programmation. Or, une utilisation naïve de cet algorithme de vérification sur des types de données mutables peut être source d'incohérences, lorsque le partage d'une valeur mutable lui impose d'avoir un type polymorphe, produit par anti-unification. Une solution évidente serait d'interdire de telles anti-unifications en faisant échouer la vérification lorsque celle-ci est amenée à anti-unifier deux termes partageant un même symbole de tête (constructeur d'un type mutable) dont les arguments ne sont pas identiques. Par exemple, si Ref est un type paramétré dont les valeurs sont des enregistrements avec un champ mutable, et que l'algorithme est amené à anti-unifier $\mathtt{Ref}(\bar{\sigma}_1)$ et $\mathtt{Ref}(\bar{\sigma}_2)$, il échouera si $\bar{\sigma}_1$ et $\bar{\sigma}_2$ ne sont pas identiques. De tels cas d'échec peuvent être vus comme le refus du partage potentiel d'une occurrence mutable vue par le contexte comme porteuse de valeurs de types incompatibles : comme dans le cas non mutable, cela correspond à une absence de valeur au moment de la dé-sérialisation, mais le caractère mutable ne garantit pas la pérennité de cette absence.

Nous n'avons pas abordé le problème des valeurs fonctionnelles dans cet article. OCaml permet leur sérialisation, sous la forme, essentiellement, de fermetures composées d'une adresse de code et d'un environnement. La de-sérialisation est capable de relocaliser de façon sûre l'adresse de code ; il reste donc, pour une dé-sérialisation sûre, à obtenir et vérifier le type de l'environnement. Or, ce dernier n'est généralement pas disponible : en effet le type d'une valeur fonctionnelle n'apporte en général pas d'information sur le type de son environnement. Nous envisageons dans un futur proche d'aborder ce problème.

Le cas des valeurs de type $\mathtt{TRepr}(\tau)$ mérite une attention particulière : en effet, elles peuvent elles aussi être sérialisées, et devront être dé-sérialisées de façon sûre. Par exemple, la sérialisation de la valeur $\ll\forall\alpha.\mathtt{List}(\alpha)\gg$ doit pouvoir être relue comme étant de type $\mathtt{TRepr(List(Int))}$ ou $\forall\alpha.\mathtt{TRepr(List}(\alpha))$, mais pas $\forall\alpha.\mathtt{TRepr}(\alpha)$. Il apparaît donc souhaitable d'autoriser la lecture de la représentation d'un type comme étant du type de la représentation d'une de ses instances.

Le type TRepr étant prédéfini, il est aisé d'étendre l'algorithme *Check* par une analyse explicite des différents cas constituant le corps de sa définition, et faisant en sorte qu'il accepte la représentation de $\top$ comme étant du type de la représentation d'un type quelconque.

10. Conclusion

Nous avons présenté dans cet article une méthode originale de vérification de types de valeurs dé-sérialisées, et nous en avons prouvé la correction. Une implantation prototype en a démontré l'effectivité, et nous envisageons l'extension de la méthode aux données mutables et fonctionnelles.

Remerciements

Nous tenons à remercier Damien Doligez et Didier Rémy pour les discussions que nous avons pu avoir avec eux à différentes étapes de ce travail, ainsi que les rapporteurs dont les commentaires nous ont incité à clarifier certains passages de cet article.

Références

[1] S. Aditya et A. Caro. « Compiler-directed Type Reconstruction for Polymorphic Languages ». Dans *Functional Programming and Computer Architecture*, pages 74–82, 1993.

[2] A. W. Appel. « Runtime Tags Aren't Necessary ». *Lisp and Symbolic Computation*, 2(2):153–162, 1989.

[3] L. Cardelli. « Amber ». Dans G. Cousineau, P.-L. Curien et B. Robinet, éditeurs, *Combinators and Functional Programming Languages*, volume 242 de *Lecture Notes in Computer Science*, pages 21–47. Springer, 1985.

[4] E. Chailloux, P. Manoury et B. Pagano. « Types behind the mirror: a proposal for partial ML type reconstruction at runtime ». Dans *Types in compilation*, 1997.

[5] R. Di Cosmo. « Deciding Type Isomorphisms in a Type-Assignment Framework ». *Journal of Functional Programming*, 3(4):485–525, 1993.

[6] K. Crary, S. Weirich et G. Morrisett. « Intensional Polymorphism in Type-Erasure Semantics ». Dans *International Conference on Functional Programming*, pages 301–312, 1998.

[7] C. Dubois, F. Rouaix et P. Weis. « Generic Polymorphism ». Dans *Principles of Programming Languages*, pages 118–129, 1995.

[8] J. Furuse et P. Weis. « Entrées-sorties de valeurs en Caml ». Dans *Actes des Journées francophones des langages applicatifs*. INRIA, janvier 2000.

[9] B. Goldberg et M. Gloger. « Polymorphic Type Reconstruction for Garbage Collection Without Tags ». Dans *LISP and Functional Programming*, pages 53–65, 1992.

[10] G. Henry. « Vers un typage sûr de la dé-sérialisation de valeurs ML ». Rapport de stage de Master MPRI, Université Pierre et Marie Curie, 2005. `http://mpri.master.univ-paris7.fr/stages-2005-rapports.html`.

[11] M. P. Herlihy et B. Liskov. « A Value Transmission Method for Abstract Data Types ». *ACM Trans. Program. Lang. Syst.*, 4(4):527–551, 1982.

[12] M. Hicks, S. Weirich et K. Crary. « Safe and flexible dynamic linking of native code ». Dans R. Harper, éditeur, *Third International Workshop on Types in Compilation*, volume 2071 de *Lecture Notes in Computer Science*, pages 147–176. Springer-Verlag, 2000.

[13] G. Huet. « *Résolution d'équations dans des langages d'ordre 1,2,..., ω* ». Thèse d'état, Université Paris 7, 1976.

[14] A. J. Kfoury, J. Tiuryn et P. Urzyczyn. « Type reconstruction in the presence of polymorphic recursion ». *ACM Trans. Program. Lang. Syst.*, 15(2):290–311, 1993.

[15] J. J. Leifer, G. Peskine, P. Sewell et K. Wansbrough. « Global abstraction-safe marshalling with hash types ». Dans *International Conference of Functional Programming*, 2003.

[16] X. Leroy, D. Doligez, J. Garrigue, D. Rémy et J. Vouillon. « *The Objective Caml system, release 3.08* ». INRIA-Rocquencourt, juillet 2004. `http://caml.inria.fr/pub/docs/manual-ocaml/`.

[17] X. Leroy et M. Mauny. « Dynamics in ML ». *Journal of Functional Programming*, 3(4):431–463, 1993.

[18] B. H. Liskov, R. A. Atkinson, T. Bloom, J. E. Moss, J. C. Schaffaert, R. W. Scheifler et A. Snyder. CLU Reference Manual. Dans *Lecture Notes in Computer Science*. Springer-Verlag, 1981.

[19] « Merriam-Webster OnLine ». `http://www.m-w.com/dictionary.htm`.

[20] A. Mycroft. « Polymorphic Type Schemes and Recursive Definitions ». Dans *Proceedings of the 6th Colloquium on International Symposium on Programming*, pages 217–228, London, UK, 1984. Springer-Verlag.

[21] G. Plotkin. « *Automatic Methods of Inductive Inference* ». PhD thesis, University of Edinburgh, 1971.

[22] Inc. Sun Microsystems. « Java Object Serialization Specification », 1998.

[23] R. E. Tarjan. « Depth first search and linear graph algorithms ». *SIAM Journal on Computing*, 1(2):146–160, 1972.

[24] A. P. Tolmach. « Tag-Free Garbage Collection Using Explicit Type Parameters ». Dans *LISP and Functional Programming*, pages 1–11, 1994.