Flexible Types

Robust type inference for first-class polymorphism

Daan Leijen

Microsoft Research

daan@microsoft.com

Abstract

We present HML, a type inference system that supports full first-class polymorphism where few annotations are needed: only function parameters with a polymorphic type need to be annotated. HML is a simplification of MLF where only flexibly quantified types are used. This makes the types easier to work with from a programmers perspective, and simplifies the implementation of the type inference algorithm. Still, HML retains much of the expressiveness of MLF, it is robust with respect to small program transformations, and has a simple specification of the type rules with an effective type inference algorithm that infers principal types. A small reference implementation with many examples is available at: `http://research.microsoft.com/users/daan/pubs.html`.

Categories and Subject Descriptors D.3.3 [*Programming Languages*]: Language Constructs and Features—Polymorphism

General Terms Languages, Design, Theory

Keywords First-class polymorphism, System F, MLF

1. Introduction

Most type inference systems used in practice are based on Hindley-Milner type inference (Hindley 1969; Milner 1978; Damas and Milner 1982). This is an attractive type system since it has a simple logical specification, and an effective type inference algorithm that can automatically infer most general, or *principal*, types for expressions without any further type annotations. Unfortunately, Hindley-Milner inference does not support first-class polymorphic values, and only allows a simple form of polymorphism on let-bound values. This is a severe restriction in practice. Even though uses of first-class polymorphism occur infrequently, there is usually no good alternative or work around (see (Peyton Jones et al. 2007) for a good overview).

The reference calculus for first-class polymorphism is System F which is explicitly typed. As remarked by Rémy (2005) one would like to have the expressiveness of System F combined with the convenience of Hindley-Milner type inference. Unfortunately, full type inference for System F is undecidable (Wells 1999). Therefore, the only way to achieve our goal is to augment Hindley-Milner type inference with just enough programmer provided annotations to make programming with first-class polymorphism practical.

There is a long list of research papers that propose type inference systems for System F (Peyton Jones et al. 2007; Rémy 2005; Jones 1997; Le Botlan and Rémy 2003; Le Botlan 2004; Odersky and Läufer 1996; Garrigue and Rémy 1999; Vytiniotis et al. 2006; Dijkstra 2005; Leijen 2008a). All of these systems mostly differ in where type annotations are needed in the program, and how easy it is for the programmer to determine where to put them in practice.

The MLF type inference system (Le Botlan and Rémy 2003; Le Botlan 2004) is the most expressive inference system to date, and requires type annotations on function parameters that are used at two or more polymorphic types. Despite its good properties, MLF has found little acceptance in practice, perhaps partly due to the intricate type structure with both flexible and rigid quantification, and a more involved type inference algorithm.

In this article, we present a simplification of MLF, called HML, that uses only flexible types. This simplifies both the types and the inference algorithm, while retaining many of the good properties of MLF. In particular:

- HML is a simplification and restriction of MLF, and has exactly the same expressiveness of Implicit MLF (Le Botlan and Rémy 2007). A novel feature is that HML only uses flexible types and avoids rigidly quantified types completely. This makes types easier to work with from a programmers perspective, and we show that it also simplifies the type inference algorithm.
- HML has a clear annotation rule: only function parameters with a polymorphic type need to be annotated.
- As a consequence, HML is a conservative extension of Hindley-Milner: every program that is well-typed in Hindley-Milner is also a well-typed HML program and type annotations are never required for such programs.
- The type rules of HML are very similar to MLF, where flexible types enable principal derivations where every expression can be assigned a most general type. This allows a programmer to use modular reasoning about a program, and enables efficient type inference.
- HML is very robust with respect to small program transformations. It has the property that whenever the application $e_1\ e_2$ is well-typed, so is the abstraction $apply\ e_1\ e_2$ and reverse application $revapp\ e_2\ e_1$. Moreover, we can always inline a let-binding, or abstract a common expression into a shared let-binding. We consider this an important property as it forms the basis of equational reasoning where we expect to be able to replace equals by equals.

In the following section we give an overview of HML in practice. Section 4 presents the formal logical type rules of HML followed by a discussion on the type equivalence and instance relation (Sec-

POPL'09, January 18–24, 2009, Savannah, Georgia, USA.

tion 5). Section 6 shows how any System F program can be expressed in HML. In Section 7 we discuss a restriction of HML where all bound values have a regular System F type. Finally, Section 8 and Section 9 describe the unification and type inference algorithm.

2. An overview

As said in the introduction, one would like to have a language with the expressiveness of System F combined with the convenience of Hindley-Milner type inference. However, since type inference for first-class polymorphism is undecidable in general (Wells 1999) some type annotation are always needed. Previously proposed type inference systems that support first-class polymorphism mostly differ where such annotations are needed in the program, and how easy it is for the programmer to apply the those rules in practice. HML has a very clear annotation rule:

function parameters with a polymorphic type need an annotation

and that's it! In practice this means that an annotation is only needed when defining a function with polymorphic parameters. Take for example the following function:

$$poly = \lambda(f :: \forall\alpha.\, \alpha \to \alpha).\, (f\ 1, f\ True)$$

The parameter f of the $poly$ function must be annotated with a polymorphic type since it is applied to both an integer and a boolean value. In general, one can never infer a type for such parameter since there are many possible types that f can have; for example $\forall\alpha.\, \alpha \to \alpha \to \alpha$ or $\forall\alpha.\, \alpha \to Int$. Therefore HML simply requires annotations on all parameters with a polymorphic type.

These are the only annotations that are ever required, and polymorphic functions and data structures can be used freely without further annotations. Assuming the following definitions:

$$\begin{array}{ll} id & :: \forall\alpha.\, \alpha \to \alpha \\ apply & :: \forall\alpha\beta.\, (\alpha \to \beta) \to \alpha \to \beta \\ revapp & :: \forall\alpha\beta.\, \alpha \to (\alpha \to \beta) \to \beta \end{array}$$

HML can type $poly\ \ id$ or $poly\ \ (\lambda x.\, x)$ for example. Impredicative instantiations are also inferred automatically. For example $apply\ poly\ id$, or $revapp\ id\ poly$, are accepted where the α quantifier of $apply$ and $revapp$ is instantiated to the polymorphic type $\forall\alpha.\, \alpha \to \alpha$. In general, HML is not sensitive to the order of applications and whenever the application $e_1\ \ e_2$ is accepted, so is $apply\ \ e_1\ \ e_2$ and $revapp\ \ e_2\ \ e_1$. This is an important property in practice since it allows us to apply the usual abstractions uniformly over polymorphic values too.

2.1 Flexible types

Type inference works well with data structures with polymorphic elements too. Assuming:

$$\begin{array}{ll} inc & :: Int \to Int \\ single & :: \forall\alpha.\, \alpha \to List\, \alpha \\ append & :: \forall\alpha.\, List\, \alpha \to List\, \alpha \to List\, \alpha \\ map & :: \forall\alpha\beta.\, (\alpha \to \beta) \to List\, \alpha \to List\, \beta \end{array}$$

we can for example map the $poly$ function over a list of polymorphic identity functions as $map\ poly\ (single\ id)$, where $single\ id$ has type $List\, (\forall\alpha.\, \alpha \to \alpha)$. Of course, sometimes a monomorphic type is inferred the expression $single\ \ id$. Take for example $append\ \ (single\ \ inc)\ \ (single\ \ id)$ where $single\ \ id$ must get the type $List\, (Int \to Int)$. We can wonder now what happens when we share the $single\ id$ expression, as in the following program:

$$\begin{array}{l} \mathbf{let}\ ids = single\ id \\ \mathbf{in}\ \ (map\ poly\ ids, append\ (single\ inc)\ ids) \end{array}$$

Of course, according to our annotation rule, this program is accepted as is in HML. This implies that we must be able to use ids both as a list of polymorphic elements, $List\ (\forall\alpha.\, \alpha \to \alpha)$, and as a list of integer functions, $List\ \ (Int \to Int)$, even though these types are incomparable in System F.

To address this, HML uses *flexible types* to assign a most general type to ids, namely $\forall(\beta \geqslant \forall\alpha.\, \alpha \to \alpha).\, List\, \beta$. We can read this as "for any type β that is an instance of $\forall\alpha.\, \alpha \to \alpha$, this is a list of β", and we can instantiate this type to both of the above types, i.e.:

$$\begin{array}{lcl} \forall(\beta \geqslant \forall\alpha.\, \alpha \to \alpha).\, List\, \beta & \sqsubseteq & List\, (\forall\alpha.\, \alpha \to \alpha) \\ \forall(\beta \geqslant \forall\alpha.\, \alpha \to \alpha).\, List\, \beta & \sqsubseteq & \forall\alpha.\, List\, (\alpha \to \alpha) \\ \forall(\beta \geqslant \forall\alpha.\, \alpha \to \alpha).\, List\, \beta & \sqsubseteq & List\, (Int \to Int) \\ \ldots & & \end{array}$$

Flexible types are key to enable modular type inference where every expression can be assigned a most general, or principal, type. By putting the bound on the quantifier, flexible types also neatly keep track shared polymorphic types. Take for example the $choose$ function:

$$choose :: \forall\alpha.\, \alpha \to \alpha \to \alpha$$

The expression $choose\ \ id$ can be assigned two incomparable types in System F, namely $\forall\alpha.\, (\alpha \to \alpha) \to \alpha \to \alpha$ or $(\forall\alpha.\, \alpha \to \alpha) \to (\forall\alpha.\, \alpha \to \alpha)$. In HML, we can assign a principal type to this expression, namely $\forall(\beta \geqslant \forall\alpha.\, \alpha \to \alpha).\, \beta \to \beta$ that can be instantiated to both of the above System F types.

2.2 Robustness

Flexible types make the system very robust under rewrites. We have seen for example the system is insensitive to the order of arguments: whenever $e_1\ \ e_2$ is accepted, so is $apply\ \ e_1\ \ e_2$ and $revapp\ \ e_2\ \ e_1$. Also, we can always abstract or inline a shared expression: whenever $\mathbf{let}\ \ x\ =\ e_1\ \ \mathbf{in}\ \ e_2$ is accepted, so is $[x := e_1]e_2$, and the other way around, where we share a common expression through a let-binding. We consider this an important property as it stands at the basis of equational reasoning where equals can be substituted for equals.

The system is not entirely robust with regard to η-expansion though since all polymorphic parameters must be annotated. For example we cannot η-expand $poly$ to $\lambda f.\, poly\ \ f$ since f has a polymorphic type, and we need to write $\lambda(f :: \forall\alpha.\, \alpha \to \alpha).\, poly\, f$.

In HML, we only allow flexible types on let-bindings and cannot pass values of a flexible type as an argument. Therefore, we cannot replace let-bindings with a flexible type by lambda abstractions. This situation is similar in Hindley-Milner where only let-bindings can have a quantified type scheme and where lambda bindings are restricted to monomorphic types. Similarly, in HML only let-bindings can have a flexibly quantified type scheme and lambda bindings are restricted to regular System F types.

2.3 Implicit MLF

Implicit MLF (*i*MLF) is a restriction of MLF where flexible types cannot be assigned to function parameters (Le Botlan and Rémy 2007). This restriction was introduced to make it possible to assign a semantic meaning to flexible types in terms of System F types. Implicit MLF has no type annotations at all though and type inference is undecidable for this system. Le Botlan and Rémy describe a decidable variant called Explicit MLF (*e*MLF) that needs annotations on all function parameters that are *used* at two or more polymorphic instances. Unfortunately, this annotation rule leads to the introduction of rigidly quantified types that complicate the type system and its inference algorithm (but does support η-expansion).

HML can be seen as another decidable variant of Implicit MLF, and has exactly the same expressiveness. The key difference is a slightly more conservative annotation rule than *e*MLF where we re-

Types	σ	$::= \forall\alpha.\,\sigma$
		$\mid \alpha$
		$\mid c\ \sigma_1\ ...\ \sigma_n$
Type schemes	φ	$::= \forall(\alpha \geqslant \varphi_1).\,\varphi_2$
		$\mid \sigma$
		$\mid \bot$
Prefix	Q	$::= \alpha_1 \geqslant \hat{\varphi}_1, ..., \alpha_n \geqslant \hat{\varphi}_n$
Mono types	τ	$::= \alpha \mid c\ \tau_1\ ...\ \tau_n$
Unquantified types	ρ	$::= \alpha \mid c\ \sigma_1\ ...\ \sigma_n$
Quantified schemes	$\hat{\varphi}$	$::= \forall(\alpha \geqslant \hat{\varphi}_1).\,\varphi_2$ with $\alpha \in \mathit{ftv}(\varphi_2)$
		$\mid \bot$
Syntactic sugar	$\forall\alpha$	$= \forall(\alpha \geqslant \bot)$
	$\forall Q.\,\varphi$	$= \forall(\alpha_1 \geqslant \hat{\varphi}_1).\ ...\ \forall(\alpha_n \geqslant \hat{\varphi}_n).\,\varphi$

Figure 1. Types.

VAR
$$\frac{x : \varphi \in \Gamma}{Q, \Gamma \vdash x : \varphi}$$

INST
$$\frac{Q, \Gamma \vdash e : \varphi_1 \quad Q \vdash \varphi_1 \sqsubseteq \varphi_2}{Q, \Gamma \vdash e : \varphi_2}$$

GEN
$$\frac{(Q, \alpha \geqslant \hat{\varphi}_1), \Gamma \vdash e : \varphi_2 \quad \alpha \notin \mathit{ftv}(\Gamma)}{Q, \Gamma \vdash e : \forall(\alpha \geqslant \hat{\varphi}_1).\,\varphi_2}$$

APP
$$\frac{Q, \Gamma \vdash e_1 : \sigma_2 \to \sigma \quad Q, \Gamma \vdash e_2 : \sigma_2}{Q, \Gamma \vdash e_1\ e_2 : \sigma}$$

LET
$$\frac{Q, \Gamma \vdash e_1 : \varphi_1 \quad Q, (\Gamma, x : \varphi_1) \vdash e_2 : \varphi_2}{Q, \Gamma \vdash \ \mathbf{let}\ x = e_1\ \mathbf{in}\ e_2 : \varphi_2}$$

FUN
$$\frac{Q, (\Gamma, x : \tau) \vdash e : \sigma}{Q, \Gamma \vdash \lambda x.\, e : \tau \to \sigma}$$

FUN-ANN
$$\frac{Q, (\Gamma, x : \sigma_1) \vdash e : \sigma_2}{Q, \Gamma \vdash \lambda(x :: \sigma_1).\, e : \sigma_1 \to \sigma_2}$$

Figure 2. Type rules.

quire that *all* polymorphic function parameters must be annotated. Surprisingly, this simplifies the type system significantly since we found that rigid bounds are now no longer required, and as a result, we can also simplify the type inference algorithm where we no longer need to compute polynomial weights over types.

In Section 7 we describe a restriction of HML, called $\mathrm{HML}^{\mathsf{F}}$, which disallows flexible types on let-bindings too, and requires annotations on some let-bindings with a higher-rank type. Even though more annotations are needed, this could perhaps be a useful simplification in practice as it removes most flexible types from the realm of the programmer.

3. Types

The types of HML are equivalent to those of Implicit MLF and are defined formally in Figure 1. Basic types σ are equivalent to regular System F types and are either a type variable α, a constructor application $c\ \sigma_1\ ...\ \sigma_n$, or a quantified type $\forall\alpha.\,\sigma$. We do not treat the function constructor $\to$ specially and assume it is part of the constructors c.

Lambda bound values always have a σ type. In contrast, let bound values can have a flexible type or *type scheme* φ. Type schemes are either a System F type σ, the most polymorphic type bottom ($\bot$), or a quantified type $\forall(\alpha \geqslant \varphi_1).\,\varphi_2$.

A quantifier with a flexible bound can be instantiated to any instance of its bound. In particular, a quantifier $\forall(\alpha \geqslant \bot)$ can be instantiated to any type since $\bot$ is the most polymorphic type. We call such bound an *unconstrained bound* and usually shorten it to $\forall\alpha$. As we see later, the $\bot$ type can be considered equivalent to System F type $\forall\alpha.\,\alpha$. Note that in System F, all quantifiers are always unconstrained where the bound φ is always $\bot$.

Since arbitrary types can occur in bounds, we can write types as $\forall(\alpha \geqslant \mathit{Int}).\,\alpha \to \alpha$ or $\forall(\beta \geqslant \mathit{List}\ (\forall\alpha.\,\alpha \to \alpha)).\,\beta$ where the bounds cannot be further instantiated. As we see later, such trivial bounds can always be removed, and the previous types are equivalent to $\mathit{Int} \to \mathit{Int}$ and $\mathit{List}\ (\forall\alpha.\,\alpha \to \alpha)$ respectively.

A *prefix* Q is a list of quantifiers $\alpha_1 \geqslant \hat{\varphi}_1, ..., \alpha_n \geqslant \hat{\varphi}_n$, where we assume that all the quantified type variables are distinct and form the domain of Q, written as $dom(Q)$, e.g. the set $\{\alpha_1, ..., \alpha_n\}$. In contrast to types, we do not allow any trivial bounds inside a prefix but restrict the bounds to non-trivially *quantified types* which we write with a 'hat' symbol ($\hat{\varphi}$). For example, trivial bounds like $\alpha \geqslant \mathit{Int}$ or $\alpha \geqslant \forall\beta.\,\mathit{Int}$ are disallowed inside a prefix.

We define *monomorphic types* τ as equivalent to Hindley-Milner monomorphic types, and *unquantified types* ρ as types without outer quantifiers. The free type variables of a type are defined in the usual way, where we take special care for types appearing in a bound:

$$\begin{array}{ll}
\mathit{ftv}(\bot) & = \varnothing \\
\mathit{ftv}(\alpha) & = \{\alpha\} \\
\mathit{ftv}(c\ \sigma_1\ ...\ \sigma_n) & = \mathit{ftv}(\sigma_1) \cup\ ...\ \cup \mathit{ftv}(\sigma_n) \\
\mathit{ftv}(\forall(\alpha \geqslant \varphi_1).\,\varphi_2) & \\
\quad = \mathit{ftv}(\varphi_1) \cup (\mathit{ftv}(\varphi_2) - \alpha) & \text{iff } \alpha \in \mathit{ftv}(\varphi_2) \\
\quad = \mathit{ftv}(\varphi_2) & \text{otherwise}
\end{array}$$

and this definition is extended in the natural way for sets, environments, and other structures containing types.

4. Type rules

The type rules of HML are defined in Figure 2. The expression $Q, \Gamma \vdash e : \varphi$ states that an expression e can be assigned a type φ under a type environment Γ and prefix Q. The type environment Γ binds variables to types where we write $\Gamma, x : \sigma$ to extend the environment Γ with a new binding x with type σ (replacing any previous binding for x). The prefix Q contains all the (flexible) bounds of the free type variables in Γ, e, and φ, and we have $(\mathit{ftv}(\Gamma) \cup \mathit{ftv}(e) \cup \mathit{ftv}(\varphi)) \subseteq dom(Q)$. In the definition of the Hindley-Milner type rules, the prefix Q always contains unconstrained bounds $(\alpha \geqslant \bot)$ and is therefore usually left implicit. The expression language is standard and defined as:

$$\begin{array}{lll}
e ::= & x & \text{(variable)} \\
& e_1\ e_2 & \text{(application)} \\
& \lambda x.\, e & \text{(lambda expression)} \\
& \lambda(x :: \sigma).\, e & \text{(annotated lambda expression)} \\
& \mathbf{let}\ x = e_1\ \mathbf{in}\ e_2 & \text{(let-binding)}
\end{array}$$

Most rules are almost equivalent to the usual Hindley-Milner rules, except that all of them are stated under the prefix Q. The variable rule VAR assigns the type that is found in the environment to a variable. The instance rule INST states that we can always use an instance of a derived type, where we write $Q \vdash \varphi_1 \sqsubseteq \varphi_2$ to denote that φ_2 is an instance of φ_1 under a prefix Q. The generalization rule GEN moves an assumed bounded type from the prefix to the type, effectively generalizing over that type variable. The application rule APP requires the argument and parameter type to be equivalent. Note that since the parameter type occurs under the arrow, the types in an application are σ types (and not type schemes). In contrast, the rule for let-bindings binds a generalized type scheme φ to the binding variable.

The rule for lambda expressions FUN is key to the HML system. Like Hindley-Milner, it restricts the type of the parameter to monomorphic types τ only. As we saw in the introduction, this is essential to avoid guessing polymorphic types. Note that it is still possible to derive a type for a parameter that has free type variables with a flexible bound. For example, for the abstraction $\lambda x.\ \mathit{choose}\ \mathit{id}\ x$, we can derive the principal type $\forall(\beta \geqslant \forall\alpha.\ \alpha \to \alpha).\ \beta \to \beta$ without further annotations.

At this point the reader may worry that it is possible to 'hide' a polymorphic type in the prefix to derive polymorphic type for a parameter but such is not the case: In particular, the invariant that all bounds in Q are non-trivially quantified types $\hat{\varphi}$, ensures that we can never use the parameter x in a polymorphic way. We discuss this in more detail in Section 5.4

In order to pass polymorphic parameters, the rule FUN-ANN must be used where the annotation is restricted to regular System F types σ. In this rule, the annotated (polymorphic) type of the parameter is assumed in the environment. A nice property of HML is that System F annotations on expressions $e :: \sigma$ can be encoded as an application to an annotated identity function $(\lambda(x :: \sigma).\ x)\ e$. General annotations on expressions would include flexible types though, and in practice we can add the following rule for expression annotations:

$$\textsc{Ann}\quad \frac{Q, \Gamma \vdash e : \varphi}{Q, \Gamma \vdash (e :: \varphi) : \varphi}$$

Of course, expression annotations are never necessary in a typing derivation (as implied by our annotation rule in the introduction).

4.1 Some example derivations

As an example of a complete derivation using these type rules, we derive the type for the expression $\mathit{single}\ \mathit{id}$ under some initial prefix Q_0 and environment Γ. First, we show how we can directly derive the type $\mathit{List}\,(\forall\alpha.\ \alpha \to \alpha)$ for $\mathit{single}\ \mathit{id}$, using σ_{id} as a shorthand for $\forall\alpha.\ \alpha \to \alpha$:

$$\frac{\dfrac{\begin{array}{c}Q_0, \Gamma \vdash \mathit{single} : \forall\alpha.\ \alpha \to \mathit{List}\,\alpha \\ Q_0 \vdash \forall\alpha.\ \alpha \to \mathit{List}\,\alpha \sqsubseteq (\sigma_{id} \to \mathit{List}\,\sigma_{id})\end{array}}{Q_0, \Gamma \vdash \mathit{single} : \sigma_{id} \to \mathit{List}\,\sigma_{id}} \quad Q_0, \Gamma \vdash \mathit{id} : \sigma_{id}}{Q_0, \Gamma \vdash \mathit{single}\ \mathit{id} : \mathit{List}\,(\forall\alpha.\ \alpha \to \alpha)}$$

Note that the instance relation includes the normal System F instance relation and we can immediately instantiate the type of single to $(\forall\alpha.\ \alpha \to \alpha) \to \mathit{List}\,(\forall\alpha.\ \alpha \to \alpha)$. Of course, as shown in the introduction, the most general type in HML for $\mathit{single}\ \mathit{id}$ is $\forall(\beta \geqslant \forall\alpha.\ \alpha \to \alpha).\ \mathit{List}\,\beta$, and we can derive this type as follows, using Q_1 as a shorthand for $(Q_0, \beta \geqslant \forall\alpha.\ \alpha \to \alpha)$:

$$\frac{\dfrac{\dfrac{\begin{array}{c}Q_1, \Gamma \vdash \mathit{single} : \forall\alpha.\ \alpha \to \mathit{List}\,\alpha \\ Q_1 \vdash \forall\alpha.\ \alpha \to \mathit{List}\,\alpha \sqsubseteq \beta \to \mathit{List}\,\beta\end{array}}{Q_1, \Gamma \vdash \mathit{single} : \beta \to \mathit{List}\,\beta} \quad \dfrac{\begin{array}{c}Q_1, \Gamma \vdash \mathit{id} : \forall\alpha.\ \alpha \to \alpha \\ Q_1 \vdash \forall\alpha.\ \alpha \to \alpha \sqsubseteq \beta\end{array}}{Q_1, \Gamma \vdash \mathit{id} : \beta}}{Q_1, \Gamma \vdash \mathit{single}\ \mathit{id} : \mathit{List}\,\beta \quad \beta \notin \mathit{ftv}(\Gamma)}}{Q_0, \Gamma \vdash \mathit{single}\ \mathit{id} : \forall(\beta \geqslant \forall\alpha.\ \alpha \to \alpha).\ \mathit{List}\,\beta}$$

Since the prefix contains the assumption $(\beta \geqslant \forall\alpha.\ \alpha \to \alpha)$, we know that β will be instance of $\forall\alpha.\ \alpha \to \alpha$ and therefore, we can safely instantiate the type of id to β (but not the other way around). It remains to make this intuition precise, we define the type instance ($\sqsubseteq$) and type equivalence relation ($\equiv$) formally in the next section.

5. Type equivalence and type instance

Before defininig the type instance ($\sqsubseteq$) and equivalence relation ($\equiv$), we first we establish some notation for substitutions.

5.1 Substitution

A substitution θ is a function that maps type variables to types. The empty substitution is the identity function and written as $[\,]$. We write θx for the application of a substitution θ to x where only the free type variables in x are substituted. We often write a substitution as a finite map $[\alpha_1 := \sigma_1, ..., \alpha_n := \sigma_n]$ (also written as $[\overline{\alpha} := \overline{\sigma}]$) which maps α_i to σ_i and all other type variables to themselves. The domain of a substitution contains all type variables that map to a different type: $\mathit{dom}(\theta) = \{\alpha \mid \theta\alpha \neq \alpha\}$. The codomain is a set of types and defined as: $\mathit{codom}(\theta) = \{\theta\alpha \mid \alpha \in \mathit{dom}(\theta)\}$. We write $(\alpha := \sigma) \in \theta$ if $\alpha \in \mathit{dom}(\theta)$ and $\theta\alpha = \sigma$. The expression $(\theta - \overline{\alpha})$ removes $\overline{\alpha}$ from the domain of θ, i.e. $(\theta - \overline{\alpha}) = [\alpha := \sigma \mid (\alpha := \sigma) \in \theta \wedge \alpha \notin \overline{\alpha}]$. Finally, we only consider *idempotent* substitutions θ where $\theta(\theta x) = \theta x$ (and therefore $\mathit{ftv}(\mathit{codom}(\theta)) \# \mathit{dom}(\theta)$).

5.2 A semantic definition of type equivalence and instance

The equivalence relation defines when types are considered equal and abstracts from syntactical artifacts like unbound quantifiers or the order of quantifiers. The instance relation is an extension of the System F instance relation that takes flexible bounds into account. The standard System F instance relation ($\sqsubseteq_{\mathsf{F}}$) is defined as:

$$\frac{\overline{\beta} \# \mathit{ftv}(\forall\overline{\alpha}.\ \sigma_1)}{\forall\overline{\alpha}.\ \sigma_1 \sqsubseteq_{\mathsf{F}} \forall\overline{\beta}.\ [\overline{\alpha} := \overline{\sigma}]\sigma_1}$$

where we write ($\#$) for disjoint sets. Note that only outer quantifiers can be instantiated, for example:

$$\begin{array}{lll}\forall\alpha.\ \alpha \to \alpha & \sqsubseteq_{\mathsf{F}} & \mathit{Int} \to \mathit{Int} \\ \forall\alpha.\ \alpha \to \alpha & \sqsubseteq_{\mathsf{F}} & \forall\beta.\ \mathit{List}\,(\forall\alpha.\ \alpha \to \beta) \to \mathit{List}\,(\forall\alpha.\ \alpha \to \beta)\end{array}$$

To extend this definition to flexible types, we are going to interpret flexible types as *sets of System F types*. We can naturally interpret a type $\forall(\alpha \geqslant \sigma_1).\ \sigma_2$ as all instances of type σ_2 where α is an instance of σ_1. We write $[\![\varphi]\!]$ for the semantics of φ as the set of System F types that are instances of φ. For example, $[\![\forall(\beta \geqslant \forall\alpha.\ \alpha \to \alpha).\ \mathit{List}\,\beta]\!]$ is the set of types:

$$\{\,\mathit{List}\,(\forall\alpha.\ \alpha \to \alpha), \forall\beta.\ \mathit{List}\,(\beta \to \beta), \mathit{List}\,(\mathit{Int} \to \mathit{Int}), ...\}$$

Following the approach in (Le Botlan and Rémy 2007) we define the semantics of types formally as:

Definition 1 (*Semantics of types*): The semantics of a type φ, written as $[\![\varphi]\!]$, is the set of its System F instances defined as:

$$\begin{array}{ll}[\![\bot]\!] & = [\![\forall\alpha.\ \alpha]\!] \\ [\![\sigma]\!] & = \{\sigma' \mid \sigma \sqsubseteq_{\mathsf{F}} \sigma'\} \\ [\![\forall(\alpha \geqslant \varphi_1).\ \varphi_2]\!] & = \{\sigma' \mid \sigma_1 \in [\![\varphi_1]\!], \sigma_2 \in [\![\varphi_2]\!], \\ & \qquad \overline{\beta} \# \mathit{ftv}(\forall(\alpha \geqslant \varphi_1).\ \varphi_2), \\ & \qquad \sigma' \in [\![\forall\overline{\beta}.\ [\alpha := \sigma_1]\sigma_2]\!]\}\end{array}$$

Types are considered equivalent whenever their semantics are equal:

Definition 2 (*Type equivalence*): We write $\varphi_1 \equiv \varphi_2$ for equivalence between types. It holds if and only if $[\![\varphi_1]\!] = [\![\varphi_2]\!]$.

In contrast to (Implicit) MLF we do not define equivalence under a prefix which is a consequence of restricting the bounds in a prefix

$$\textsc{Eq-Var} \quad \forall(\alpha \geqslant \varphi).\, \alpha \equiv \varphi$$

$$\textsc{Eq-Subst} \quad \forall(\alpha \geqslant \rho).\, \varphi \equiv [\alpha := \rho]\varphi$$

$$\textsc{Eq-Free} \quad \frac{\alpha \notin \mathit{ftv}(\varphi_2)}{\forall(\alpha \geqslant \varphi_1).\, \varphi_2 \equiv \varphi_2}$$

$$\textsc{Eq-Context} \quad \frac{\varphi_1 \equiv \varphi_2}{\forall(\alpha \geqslant \varphi_1).\, \varphi \equiv \forall(\alpha \geqslant \varphi_2).\, \varphi}$$

$$\textsc{Eq-Prefix} \quad \frac{\varphi_1 \equiv \varphi_2}{\forall(\alpha \geqslant \varphi).\, \varphi_1 \equiv \forall(\alpha \geqslant \varphi).\, \varphi_2}$$

$$\textsc{Eq-Comm} \quad \frac{\alpha_1 \neq \alpha_2 \quad \alpha_1 \notin \mathit{ftv}(\varphi_2) \quad \alpha_2 \notin \mathit{ftv}(\varphi_1)}{\forall(\alpha_1 \geqslant \varphi_1)(\alpha_2 \geqslant \varphi_2).\, \varphi \equiv \forall(\alpha_2 \geqslant \varphi_2)(\alpha_1 \geqslant \varphi_1).\, \varphi}$$

Figure 3. Syntactic type equivalence.

$$\textsc{I-Bottom} \quad Q \vdash \bot \sqsubseteq \varphi$$

$$\textsc{I-Equiv} \quad \frac{\varphi_1 \equiv \varphi_2}{Q \vdash \varphi_1 \sqsubseteq \varphi_2}$$

$$\textsc{I-Hyp} \quad \frac{(\alpha \geqslant \hat{\varphi}) \in Q}{Q \vdash \hat{\varphi} \sqsubseteq \alpha}$$

$$\textsc{I-Subst} \quad \frac{\alpha \notin \mathit{etv}(\varphi)}{Q \vdash \forall(\alpha \geqslant \sigma).\, \varphi \sqsubseteq [\alpha := \sigma]\varphi}$$

$$\textsc{I-Context} \quad \frac{Q \vdash \varphi_1 \sqsubseteq \varphi_2}{Q \vdash \forall(\alpha \geqslant \varphi_1).\, \varphi \sqsubseteq \forall(\alpha \geqslant \varphi_2).\, \varphi}$$

$$\textsc{I-Prefix} \quad \frac{(Q, \alpha \geqslant \hat{\varphi}) \vdash \varphi_1 \sqsubseteq \varphi_2 \quad \alpha \notin \mathit{dom}(Q)}{Q \vdash \forall(\alpha \geqslant \hat{\varphi}).\, \varphi_1 \sqsubseteq \forall(\alpha \geqslant \hat{\varphi}).\, \varphi_2}$$

Figure 4. Syntactic type instance.

to non-trivial types only. As we will see in the Section 5.4 this is an essential part of HML to ensure that unannotated parameters cannot be assigned a polymorphic type.

The definition of the instance relation must still be stated under a prefix Q though. In particular, we would like to have the following hypothesis rule to hold:

$$\textsc{I-Hyp} \quad \frac{(\alpha \geqslant \hat{\varphi}) \in Q}{Q \vdash \hat{\varphi} \sqsubseteq \alpha}$$

This rule states that whenever we have the assumption that α is an instance of φ in the prefix Q, we can instantiate that type φ to α. The example type derivation of $\mathit{single}\ \mathit{id}$ in the previous section shows an example of the use of this rule.

Therefore, to define instantiation, we first give an interpretation of prefixes. A prefix Q is meant to capture all the possible types that may be substituted for the type variables in the domain of the prefix, and we interpret a prefix as a *set of substitutions*.

Definition 3 (*Semantics of prefixes*): The semantics of a prefix Q is written as $[\![Q]\!]$, and represents the set of System F substitutions that capture all possible types that can be substituted for the type variables in the domain of Q. We define this formally as:

$$\begin{array}{ll} [\![\varnothing]\!] & = \{[\,]\} \\ [\![Q, \alpha \geqslant \hat{\varphi}]\!] & = \{\theta \circ [\alpha := \sigma] \mid \theta \in [\![Q]\!], \sigma \in [\![\hat{\varphi}]\!]\} \end{array}$$

We can now define the instance relation simply as a subset relation on the set of System F instances:

Definition 4 (*Type instance*): We write $Q \vdash \varphi_1 \sqsubseteq \varphi_2$, if a type φ_2 is an instance of a type φ_1 under a prefix Q. It holds if and only if forall $\theta \in [\![Q]\!]$, we have $\theta[\![\varphi_2]\!] \subseteq \theta[\![\varphi_1]\!]$.

Usually, we write $\varphi_1 \sqsubseteq \varphi_2$ when $\varnothing \vdash \varphi_1 \sqsubseteq \varphi_2$.

5.3 A syntactic definition of equivalence and instantiation

In practice, a semantic definition is not always easy to work with, and we discuss an equivalent formulation of equivalence and instantiation based on syntactic rules that are inductive.

In particular, we can define equivalence as the smallest transitive, symmetric, and reflexive relation that satisfies the rules of Figure 3. The first two rules are the most interesting: the rule Eq-Var inlines the bound of a variable type that is directly quantified, and Eq-Subst allows the inlining of trivial ρ bounds (that can no longer be instantiated). The rules Eq-Free and Eq-Comm ensure that unbound quantifiers are irrelevant, and that quantifiers can be rearranged. The rules Eq-Context and Eq-Prefix allow us to apply equivalence aribrarily deep inside a type.

Similarly, we define the instance relation as the smallest transitive relation that satisfies the rules of Figure 4. The rule I-Equiv includes the equivalence relation. The rule I-Bottom shows that bottom can be instantiated to any type. The substitution rule I-Subst allows us to inline an instantiated bound σ. A small technicality is that the bound should not occur in the *exposed* type variables, defined as:

$$\begin{array}{ll} \mathit{etv}(\bot) & = \varnothing \\ \mathit{etv}(c\ \sigma_1 \ldots \sigma_n) & = \varnothing \\ \mathit{etv}(\alpha) & = \{\alpha\} \\ \mathit{etv}(\forall(\alpha \geqslant \varphi_1).\, \varphi_2) & = \mathit{etv}(\varphi_1) \cup (\mathit{etv}(\varphi_2) - \{\alpha\}) \end{array}$$

This condition prevents the loss of sharing through trivial bounds. For example, take the type $\forall(\alpha \geqslant \sigma).\, \alpha \to \alpha$ which is equivalent to $\forall(\alpha \geqslant \sigma).\, \forall(\beta \geqslant \alpha).\, \forall(\gamma \geqslant \alpha).\, \beta \to \gamma$. Without the condition, we would be able to conclude in error that we can instantiate this type to $\forall(\beta \geqslant \sigma).\, \forall(\gamma \geqslant \sigma).\, \beta \to \gamma$ where the original sharing is lost and where β and γ can be instantiated separately.

Finally, the rules I-Context and I-Prefix ensure that we can apply instantiation arbitrarily deep inside bounds.

Theorem 5 (*Syntactic type equivalence and instance is sound*): The semantic definitions of type equivalence and type instance satisfy all the rules in Figure 3 and Figure 4.

Conjecture 6 (*Syntactic type equivalence and instance is complete*): Any type equivalence and instance relation can be derived using the rules in Figure 3 and Figure 4.

The proof of soundness is straightforward induction on the equivalence and instance rules and similar to the proof of soundness for the *i*MLF instance relation in (Le Botlan and Rémy 2007). In the same work, Rémy and Le Botlan conjecture the completeness of the *i*MLF relation based on a proof sketch based on an intermediate graphic representation of types (Rémy and Yakobowski 2007).

5.4 Restricted equivalence

The equivalence relation in HML differs from the one of (Implicit) MLF as it is not stated under a prefix. This is an important technical distinction to make type inference decidable for HML. In particular, *i*MLF allows the equivalence relation to inline trivial bounds from the prefix:

$$\textsc{iEq-Hyp} \quad \frac{(\alpha \geqslant \varphi_2) \in Q \quad Q \vdash \varphi_2 \equiv \rho_2}{Q \vdash \varphi_1 \equiv [\alpha := \rho_2]\varphi_1}$$

$$\begin{array}{lll}
nf(\sigma) & = \sigma & \\
nf(\bot) & = \bot & \\
nf(\forall(\alpha \geqslant \varphi_1).\, \varphi_2) & = nf(\varphi_2) & \text{iff } \alpha \notin ftv(\varphi_2) \\
nf(\forall(\alpha \geqslant \varphi_1).\, \varphi_2) & = nf(\varphi_1) & \text{iff } nf(\varphi_2) = \alpha \\
nf(\forall(\alpha \geqslant \varphi_1).\, \varphi_2) & = nf([\alpha := \rho]\varphi_2) & \text{iff } nf(\varphi_1) = \rho \\
nf(\forall(\alpha \geqslant \varphi_1).\, \varphi_2) & = \forall(\alpha \geqslant nf(\varphi_1)).\, nf(\varphi_2) &
\end{array}$$

Figure 5. Normal form.

$$\begin{array}{ll}
\mathcal{F}[\![\mathsf{x}]\!] & = x \\
\mathcal{F}[\![\mathsf{e}_1\ \mathsf{e}_2]\!] & = \mathcal{F}[\![\mathsf{e}_1]\!]\ \mathcal{F}[\![\mathsf{e}_2]\!] \\
\mathcal{F}[\![\mathsf{e}\ \sigma]\!] & = \mathcal{F}[\![\mathsf{e}]\!] \\
\mathcal{F}[\![\Lambda\alpha \cdot \mathsf{e}]\!] & = \mathcal{F}[\![\mathsf{e}]\!] \\
\mathcal{F}[\![\lambda(\mathsf{x} : \tau) \cdot \mathsf{e}]\!] & = \lambda x.\, \mathcal{F}[\![\mathsf{e}]\!] \\
\mathcal{F}[\![\lambda(\mathsf{x} : \sigma) \cdot \mathsf{e}]\!] & = \lambda(x :: \sigma).\, \mathcal{F}[\![\mathsf{e}]\!]
\end{array}$$

Figure 6. Embedding of System F.

In HML this rule never applies since prefixes cannot contain trivial bounds (but always have polymorphic bound $\hat{\varphi}$ which never equals an unquantified ρ type). Otherwise one could 'hide' a polymorphic type in the prefix, like $\forall(\alpha \geqslant List\ (\forall\alpha.\, \alpha \to \alpha))$, and assign α to the type of a lambda bound parameter x. With the above equivalence rule, we could use the INST rule to conclude that the type α is equivalent to the type $List\ (\forall\alpha.\, \alpha \to \alpha)$ under such prefix, and erronously derive a polymorphic type for the parameter x. The restriction of the prefix prevents this exactly from happening. Moreover, since IEQ-HYP does not apply we can simplify the equivalence relation and there is no need to state it under a prefix.

Even though the HML equivalence relation cannot use information from the prefix, we still maintain exactly the same expressive power as Implicit MLF. Writing e^* for an expression e with all type annotations removed, we can state the following theorem:

Theorem 7 (*Expressiveness of HML*): For any derivation in Implicit MLF, $Q, \Gamma \vdash_{\mathsf{I}} e_1 : \varphi_1$, there exists an annoted expression e_2 with $e_2^* = e_1$ where there is a HML derivation $Q, \Gamma \vdash e_2 : \varphi_2$ with $Q \vdash \varphi_1 \equiv \varphi_2$. The other direction holds too: for any HML derivation $Q, \Gamma \vdash e : \varphi_1$, we have an Implicit MLF derivation $Q, \Gamma \vdash_{\mathsf{I}} e^* : \varphi_2$ with $Q \vdash \varphi_1 \equiv \varphi_2$.

As a corollary, we have that the HML type system is sound.

Proof of Theorem 7: We observe that the type rules of *i*MLF are equivalent up to annotations and the rule IEQ-HYP. First we prove that any derivation in *i*MLF can be rewritten without using IEQ-HYP. Take any derivation that assumes a trivial bound in the generalization rule:

$$\frac{(Q, \alpha \geqslant \varphi_2), \Gamma \vdash_{\mathsf{I}} e : \varphi_1 \quad \alpha \notin ftv(\Gamma)}{Q, \Gamma \vdash_{\mathsf{I}} e : \forall(\alpha \geqslant \varphi_2).\, \varphi_1}$$

where $Q \vdash \varphi_2 \equiv \rho$ for some ρ. We can now rewrite that derivation by substituting ρ for α in the entire derivation. The result type is unchanged as $\forall(\alpha \geqslant \varphi_2).\, \varphi_1$ is equivalent to $[\alpha := \rho]\varphi_1$ (using EQ-CONTEXT, EQ-SUBST and transitivity). We can now systematically check all structural rules if they are unaffected. In particular, all uses of IEQ-HYP can be removed and substituted by EQ-SUBST. We need to be careful with I-HYP as EQ-CONTEXT can be used to change the bound to an equivalent by syntactically different type, therefore, I-HYP must be replaced by I-EQUIV. As a result, we can also remove the use of the generalization rule since $(\alpha \geqslant \varphi_2)$ is now unused in the prefix.

Any derivation in Implicit MLF can therefore be rewritten without using IEQ-HYP. Given that derivation, we can annotate any lambda bound parameters that have a derived polymorphic type, and we end up with a valid HML derivation. Since the rules of HML are a subset of *i*MLF, any derivation in HML is also valid in *i*MLF when all annotation are removed (and therefore HML is sound). □

5.5 Normal form

For type inference, it is often useful to bring types into *normal form* which makes it easier to compare them. Figure 5 defines the function $nf(\varphi)$ that returns the normal form of a type φ.

The first two cases state that F-types and $\bot$ are already in normal form. The next two cases deal with trivial bounds: unbound quantifiers are discarded (EQ-FREE), variable types are removed (EQ-VAR), and trivial bounds are inlined (EQ-SUBST). We have the following useful properties for normal forms:

Properties 8
i. $nf(\varphi) \equiv \varphi$.
ii. $nf(\varphi_1) \approx nf(\varphi_2)$ if and only if $\varphi_1 \equiv \varphi_2$.
iii. $nf(\hat{\varphi}) \neq \rho$ for any ρ.

where we use $\approx$ for syntactically equal types up to the order of the quantifiers (Le Botlan 2004).

6. Embedding of System F

It is straightforward to translate any System F program into a well-typed HML program since we only require annotations on polymorphic function parameters. Figure 6 defines the function $\mathcal{F}[\![\mathsf{e}]\!]$ that translates a System F term e to a well-typed HML term. Basically, all type abstractions and applications are removed, and only annotations on lambda bindings with polymorphic types are retained. Writing $\Gamma \vdash_{\mathsf{F}} \mathsf{e} : \sigma$ for the standard System F type system, we have the following result:

Theorem 9 (*Embedding of System F*): If $\Gamma \vdash_{\mathsf{F}} \mathsf{e} : \sigma$ then we have $\varnothing, \Gamma \vdash \mathcal{F}[\![\mathsf{e}]\!] : \sigma$.

The proof is by straightforward induction on the type rules of System F, observing that the System F instance relation is included in the HML instance relation.

7. Restricting let-bindings to System F types

As remarked in the introduction we can restrict HML let-bindings to System F types only. We call this system HML^{F}. The advantage of such restriction is that programmers are less exposed to flexible types since all bindings have a System F types. Only error messages will potentially show (principal) flexible types. A drawback is of course that certain let-bindings now require an annotation.

In particular, we design HML^{F} such that it instantiates a principal flexible type of a let-binding to its most general System F type that has the least number of inner quantifiers, which ensures compatibility with standard Hindley-Milner inference.

We do this using the function $ftype(\cdot)$ defined in Figure 7 which instantiates a flexible type φ to a most general System F type σ with the least inner polymorphism possible. For example, for the binding **let** $ids = single\ id$ it will instantiate the type $\forall(\beta \geqslant \forall\alpha.\, \alpha \to \alpha).\, List\ \beta$ to the expected Hindley-Milner type $\forall\alpha.\, List\ (\alpha \to \alpha)$. In the $ftype(\cdot)$ function, the last case in particular implements this strategy where a type of the form $\forall(\alpha \geqslant \forall Q.\, \rho).\, \varphi$ is instantiated to $\forall Q.\, [\alpha := \rho]\varphi$ [1]. It is easy to check that the following properties hold:

Properties 10
i. $ftype(\sigma) = \sigma$.

[1] Another strategy could be to keep types as polymorphic as possible, tranforming $\forall(\alpha \geqslant \varphi_1).\, \varphi_2$ to $[\alpha := ftype(\varphi_1)]ftype(\varphi_2)$ but we rather stay compatible with the expected Hindley-Milner form.

$$
\begin{array}{ll}
\mathit{ftype}(\varphi) & = \mathit{ft}(\mathit{nf}(\varphi)) \\
& \text{where} \\
& \quad \begin{array}{ll}
\mathit{ft}(\sigma) & = \sigma \\
\mathit{ft}(\bot) & = \forall\alpha.\, \alpha \\
\mathit{ft}(\forall(\alpha \geqslant \bot).\, \varphi) & = \forall\alpha.\, \mathit{ft}(\varphi) \\
\mathit{ft}(\forall(\alpha \geqslant \forall Q.\, \rho).\, \varphi) & = \mathit{ft}(\forall Q.\, [\alpha := \rho]\varphi)
\end{array}
\end{array}
$$

Figure 7. Force a flexible type to a System F type.

ii. If $\mathit{ftype}(\varphi) = \sigma$ then we have $\varphi \sqsubseteq \sigma$.

Formally, HML^{F} consists of the normal HML type rules given in Figure 2, where the LET rule is replaced by the rule LET-F defined in Figure 8. The rule LET-F requires that the type φ_1 derived for e_1 is the most general type possible, i.e. forall φ_0 such that $Q, \Gamma \vdash e_1 : \varphi_0$ holds, we have that $Q \vdash \varphi_1 \sqsubseteq \varphi_0$. The restriction to most general types is needed to ensure that we still have principle type derivations. Finally, the derived type for e_1 is instantiated to a System F type using the function $\mathit{ftype}(\varphi_1)$ and bound in the environment.

A nice property of HML^{F}, is that we can give a clear (but conservative) annotation rule for let-bindings: *Only let bindings with a higher-rank type may require an annotation*. As a consequence, HML^{F} is still a conservative extension of Hindley-Milner. For example, the following expression:

$$\mathbf{let}\ ids = single\ id\ \mathbf{in}\ append\ (single\ inc)\ ids$$

is accepted without annotations where ids gets the expected Hindley-Milner type $\forall\alpha.\, \mathit{List}\ (\alpha \to \alpha)$. To construct a list of polymorphic elements though, we need to add an annotation:

$$\mathbf{let}\ ids = (single\ id :: \mathit{List}\ (\forall\alpha.\, \alpha \to \alpha))\ \mathbf{in}\ map\ poly\ ids$$

to accept this program in HML^{F}. The annotation rule is clear, but a bit conservative. For example, no annotation is needed when the most general type of a let-binding is already a plain System F type (since $\mathit{ftype}(\sigma)$ equals σ). For example, the following bindings are accepted without further annotations:

$$
\begin{array}{lll}
\mathbf{let}\ ids & = & (single\ id :: \mathit{List}\ (\forall\alpha.\, \alpha \to \alpha)) \\
\quad xs & = & cons\ id\ ids \\
\quad ys & = & tail\ ids \\
\quad zs & = & map\ id\ (tail\ (cons\ id\ ids))
\end{array}
$$

All of the above bindings have a principal type that is already a System F type ($\mathit{List}\ (\forall\alpha.\, \alpha \to \alpha)$), and no annotations are needed. Only when the most-general type of a let-binding is a flexible type, and we have a *choice* which System F types we can assign, an annotation may be required. If there is no annotation, HML^{F} always disambiguates flexible types in a predefined way that leads the System F instance with the least inner polymorphism.

Note that restricting let-bindings to System F types makes HML^{F} less expressive than HML, since we can no longer share an expression like ids at different polymorphic instances, i.e. in contrast to HML, the program:

$$
\begin{array}{l}
\mathbf{let}\ ids = single\ id \\
\mathbf{in}\ (map\ poly\ ids, append\ (single\ inc)\ ids)
\end{array}
$$

is always rejected in HML^{F} as there exists no System F type for ids to make this well typed. In practice, a language may choose to use rule LET-F if a let binding is unannotated, while still allowing the programmer to annotate such bindings with flexible types, maintaining the expressiveness of original system.

$$
\text{LET-F}\quad \frac{\begin{array}{c} Q, \Gamma \vdash e_1 : \varphi_1 \\ \forall\varphi_0.\ \text{if } Q, \Gamma \vdash e_1 : \varphi_0 \text{ then } Q \vdash \varphi_1 \sqsubseteq \varphi_0 \\ Q, (\Gamma, x : \mathit{ftype}(\varphi_1)) \vdash e_2 : \varphi_2 \end{array}}{Q, \Gamma \vdash\ \mathbf{let}\ x = e_1\ \mathbf{in}\ e_2 : \varphi_2}
$$

Figure 8. Environment restricted to System F types.

7.1 A comparison with FPH

At this point it is interesting to compare HML^{F} with the recently introduced FPH system (Vytiniotis et al. 2008). The FPH type system supports first-class polymorphism and has a similar annotation rule as HML^{F}: all polymorphic parameters must be annotated, and only let-bindings with a higher-rank type may require an annotation. FPH achieves this by distinguishing between inferred boxy types and known regular types. An advantage of this approach is that FPH does not need to require most general types at let-bindings as in HML^{F}. A drawback though is that FPH cannot always assign principal types to expressions. For example, the type of $single\ id$ can have both the (incomparable) types $\forall\alpha.\, \mathit{List}\ (\alpha \to \alpha)$ and $\mathit{List}\ (\boxed{\forall\alpha.\, \alpha \to \alpha})$ in FPH, but we cannot assign a principal type like $\forall(\beta \geqslant \forall\alpha.\, \alpha \to \alpha).\, \mathit{List}\ \beta$ as in HML^{F}.

Given the similar annotation rules, we were hoping that the expressiveness of HML^{F} would coïncide with FPH but HML^{F} accepts slightly more programs due to the following reasons:

- FPH restricts function results to unquantified ρ types. For example, if we have a function $\mathit{foo} :: (\mathit{Int} \to (\forall\alpha.\, \alpha \to \alpha)) \to \mathit{Int}$, then the application $poly\ (\lambda x.\, x)$ is accepted, but the similar application $\mathit{foo}\ (\lambda y.\, \lambda x.\, x)$ is rejected (since $(\lambda y.\, \lambda x.\, x$ gets the type $\forall\alpha\beta.\, \alpha \to \beta \to \beta$ which is not polymorphic enough). In HML^{F}, the sub-expression $\lambda y.\, \lambda x.\, x$ has the principal type $\forall\alpha.\, \forall(\beta \geqslant \forall\alpha.\, \alpha \to \alpha).\, \alpha \to \beta$ and both applications are accepted.
- FPH restricts the type of let-bindings to non-boxy types only, even if the type is unambiguous. Take for example our earlier example that is accepted in HML^{F}:

  $$\mathbf{let}\ xs = cons\ id\ ids$$

  FPH rejects this definition without annotation since the result type contains boxes (similarly, the previous ys and zs definitions are also rejected).

Therefore, even though FPH and HML^{F} are not equal, we conjecture that FPH is a pure restriction of HML^{F}, which is itself a restriction of HML. To bring these systems closer, we can restrict HML^{F} further by allowing function results with ρ types only, but unfortunately we have not yet found a satisfactory formalization for the latter of the above restrictions.

8. Unification

Unification of flexible types is more involved than the usual unification used in Hindley-Milner, but is easier than standard MLF unification since there are no rigid bounds, and we no longer have to compute polynomial weights over types (Le Botlan 2004). The unification algorithm uses two kinds of substitutions, namely a prefix Q and a normal substitution θ. In actual implementations both substitutions are usually merged and implemented directly using updateable references on type variables (see (Leijen 2008b) for a reference implementation). Again, we have as an invariant that Q only contains bounds with quantified types. Dually, the substitution θ contains only F types: if $\varphi \in \mathit{codom}(\theta)$ then $\mathit{nf}(\varphi) = \sigma$ for some σ. Before discussing the actual unification algorithm, we first look at two helper functions over prefixes defined in Figure 11.

$unify :: (Q, \sigma_1, \sigma_2) \rightarrow (Q, \theta)$
where σ_1 and σ_2 are in normal form
and $(\alpha := \varphi) \in \theta \;\Rightarrow\; nf(\varphi) = \sigma$

$unify(Q, \alpha, \alpha) =$
return $(Q,\ [\,])$

$unify(Q_1, c\ \sigma_1\ ...\ \sigma_n, c\ \sigma'_1\ ...\ \sigma'_n) =$
let $\theta_1 \quad = [\,]$
let $\theta_{i+1} = \theta'_i \circ \theta_i$
let $(Q_{i+1}, \theta'_i) = unify(Q_i, \theta_i\sigma_i, \theta_i\sigma'_i)$
return $(Q_{n+1},\ \theta_{n+1})$

$unify(Q, \alpha, \sigma)$ or
$unify(Q, \sigma, \alpha)$ with $(\alpha \geqslant \hat{\varphi}) \in Q \wedge \sigma \notin \mathcal{V}$
fail if $(\alpha \in dom(Q/\sigma))$ ('occurs' check)
let $(Q_1, \theta_1) = subsume(Q, \sigma, \hat{\varphi})$
let $(Q_2, \theta_2) = update(Q_1, \alpha := \theta_1\sigma)$
return $(Q_2,\ \theta_2 \circ \theta_1)$

$unify(Q, \alpha_1, \alpha_2)$ with $(\alpha_1 \geqslant \hat{\varphi}_1) \in Q \wedge (\alpha_2 \geqslant \hat{\varphi}_2) \in Q$
fail if $(\alpha_1 \in dom(Q/\hat{\varphi}_2) \vee \alpha_2 \in dom(Q/\hat{\varphi}_1))$
let $(Q_1, \theta_1, \varphi) = unifyScheme(Q, \hat{\varphi}_1, \hat{\varphi}_2)$
let $(Q_2, \theta_2) \quad = update(Q_1, \alpha_1 := \alpha_2)$
let $(Q_3, \theta_3) \quad = update(Q_2, \alpha_2 \geqslant \varphi)$
return $(Q_3,\ \theta_3 \circ \theta_2 \circ \theta_1)$

$unify(Q, \forall\alpha.\, \sigma_1, \forall\beta.\, \sigma_2) =$
assume c is a fresh (skolem) constant
let $(Q_1, \theta_1) = unify(Q, [\alpha := c]\sigma_1, [\beta := c]\sigma_2)$
fail if $(c \in (con(\theta_1) \cup con(Q_1)))$
return $(Q_1,\ \theta_1)$

Figure 9. Unification.

$subsume :: (Q, \sigma, \varphi) \rightarrow (Q, \theta)$
where σ and φ are in normal form

$subsume(Q, \forall\overline{\alpha}.\, \rho_1, \forall Q_2.\, \rho_2)$
assume $dom(Q) \,\#\, dom(Q_2)$ and $\overline{c}$ are fresh constants
let $(Q_1, \theta_1) \ = unify(QQ_2, [\overline{\alpha} := \overline{c}]\rho_1, \rho_2)$
let $(Q_2, Q_3) = split(Q_1, dom(Q))$
let $\theta_2 = \theta_1 - dom(Q_3)$
fail if $(\overline{c} \in (con(\theta_2) \cup con(Q_2)))$
return $(Q_2,\ \theta_2)$

$unifyScheme :: (Q, \varphi_1, \varphi_2) \rightarrow (Q, \theta, \varphi)$
where φ_1 and φ_2 are in normal form

$unifyScheme(Q, \bot, \varphi)$ or $unifyScheme(Q, \varphi, \bot)$
return $(Q,\ [\,],\ \varphi)$

$unifyScheme(Q, \forall Q_1.\, \rho_1, \forall Q_2.\, \rho_2)$
assume the domains of Q, Q_1, and Q_2 are disjoint
let $(Q_3, \theta_3) \ = unify(QQ_1Q_2, \rho_1, \rho_2)$
let $(Q_4, Q_5) = split(Q_3, dom(Q))$
return $(Q_4,\ \theta_3,\ \forall Q_5.\, \theta_3\rho_1)$

Figure 10. Subsumption and type scheme unification.

8.1 Split and update

The split function $split(Q, \overline{\alpha})$ splits a prefix in two parts Q_1 and Q_2 such that Q_1 is the only part useful to $\overline{\alpha}$, i.e. $\forall Q.\, \alpha_1 \rightarrow ... \rightarrow \alpha_n \equiv \forall Q_1.\, \alpha_1 \rightarrow ... \rightarrow \alpha_n$ with $\overline{\alpha} = \{\alpha_1, ..., \alpha_n\}$. The split function is used for example to generalize over types where the prefix must be split in one part that can be generalized, and another part that cannot. In an actual implementation, this operation can be implemented efficiently using ranked type variables similar to the usual optimization used for generalization over free variables in the environment (Kuan and MacQueen 2007; Leijen 2008b).

The update function $update(Q, \alpha \geqslant \varphi)$ updates a prefix Q that contains a binding for α with a new binding φ taking care to maintain the invariants on Q. First the prefix is split according to the free type variables in φ to ensure that these are properly scoped. If the binding is updated with an unquantified type, we extend the substitution with $[\alpha := \rho]$, otherwise we update the prefix directly with $\alpha \geqslant \varphi$. Similarly, the expression $update(Q, \alpha := \sigma)$ assigns the type σ to α, and returns a substitution $[\alpha := \sigma]$, updating the prefix Q accordingly. In an actual implementation, the $update$ operation usually just updates a reference directly, and both $split$ and $update$ are more or less artifacts of using explicit substitutions.

8.2 Unification

The unification algorithm is defined in Figure 9 and Figure 10. The function $unify(Q, \sigma_1, \sigma_2)$ takes a prefix Q and two F types σ_1 and σ_2 and returns a new prefix Q' and a substitution θ such that $Q' \vdash \theta\sigma_1 \equiv \theta\sigma_2$. The cases for equal variables and constructors are standard.

The next two cases deal with the unification of a variable α, with a binding $(\alpha \geqslant \hat{\varphi}) \in Q$, and an F type σ. First, we ensure that no infinite type is unified with the 'occurrence check' $\alpha \notin dom(Q/\hat{\varphi})$, where $dom(Q/\varphi)$ returns the *useful* domain of Q with respect to φ. In usual ML type inference, this is always equivalent to the free type variables of φ, but in our case some type variables can be referenced indirectly. For example, $dom((\gamma \geqslant \bot, \beta \geqslant \forall\delta.\, \delta \rightarrow \gamma)/(\forall\alpha.\, \alpha \rightarrow \beta))$ is $\{\beta, \gamma\}$ even though $\gamma \notin ftv(\forall\alpha.\, \alpha \rightarrow \beta)$. Formally, we say that $\alpha \in dom(Q/\varphi)$ if and only if $Q = (Q_1, \alpha \geqslant \hat{\varphi}_1, Q_2)$ and $\alpha \in ftv(\forall Q_2.\, \varphi)$.

After the occurrence check, we try to instantiate the type $\hat{\varphi}$ to the F type σ using the $subsume$ function. If that succeeds, we update the prefix with the new substitution $\alpha := \theta_1\sigma$.

The next case deals with two variables α_1 and α_2 bound to two flexible types $\hat{\varphi}_1$ and $\hat{\varphi}_2$. Again, we first do an occurence check for both variables, and if that succeeds, we call $unifyScheme$ to find the join of $\hat{\varphi}_1$ and $\hat{\varphi}_2$, i.e. the most general type φ that is an instance of both $\hat{\varphi}_1$ and $\hat{\varphi}_2$, and update the prefix afterwards with $\alpha_1 := \alpha_2$, and $\alpha_2 \geqslant \varphi$.

Finally, we have the case of two quantified types $\forall\alpha.\, \sigma_1$ and $\forall\beta.\, \sigma_2$. In this case we replace both quantifiers with a fresh (skolem) constant c and unify the remaining types. For this to work, we assume that we have ordered the quantifiers of the F types in a canonical order where each quantifier appears according to the order of occurence in the type, see (Leijen 2007, 2008a) for details. Moreover, we need to ensure that the skolem constant does not escape into the environment, and that no free type variables unify with such constant. For example, it would be incorrect to unify $\forall\alpha.\, \alpha \rightarrow \beta$ with $\forall\alpha.\, \alpha \rightarrow \alpha$. The check $c \notin (con(\theta_1) \cup con(Q_1))$ ensures that this is the case, where the function $con(\cdot)$ returns the type constants in the codomain of the substitution θ_1 or the prefix Q_1. In an implementation based on updateable references, this check can be done efficiently by checking if the original types $\forall\alpha.\, \sigma_1$ and $\forall\beta.\, \sigma_2$ do not contain a reference to c after the unification (Leijen 2008b).

(update the prefix)
$update(Q, \alpha \geqslant \varphi_2)$
  let $(Q_0, (Q_1, \alpha \geqslant \hat{\varphi}_1, Q_2)) = split(Q, ftv(\varphi_2))$
  if $(nf(\varphi_2) = \rho)$
    then return $((Q_0, Q_1, [\alpha := \rho]Q_2),\ [\alpha := \rho])$
    else return $((Q_0, Q_1, \alpha \geqslant \varphi_2, Q_2),\ [\,])$

$update(Q, \alpha := \sigma)$
  let $(Q_0, (Q_1, \alpha \geqslant \varphi, Q_2)) = split(Q, ftv(\sigma))$
  return $((Q_0, Q_1, [\alpha := \sigma]Q_2),\ [\alpha := \sigma])$

(extend a prefix
$extend(Q, \alpha \geqslant \varphi) =$
  if $(nf(\varphi) = \rho)$
    then return $(Q,\ [\alpha := \rho])$
    else return $((Q, \alpha \geqslant \varphi),\ [\,])$

(split a prefix)
$split :: (Q, \overline{\alpha}) \rightarrow (Q, Q)$

$split(\varnothing, \overline{\alpha}) =$
  return $(\varnothing,\ \varnothing)$

$split((Q, \alpha \geqslant \hat{\varphi}), \overline{\alpha}) =$ with $\alpha \in \overline{\alpha}$
  let $(Q_1, Q_2) = split(Q, ((\overline{\alpha} - \alpha) \cup ftv(\varphi)))$
  return $((Q_1, \alpha \geqslant \hat{\varphi}),\ Q_2)$

$split((Q, \alpha \geqslant \hat{\varphi}), \overline{\alpha}) =$ with $\alpha \notin \overline{\alpha}$
  let $(Q_1, Q_2) = split(Q, \overline{\alpha})$
  return $(Q_1,\ (Q_2, \alpha \geqslant \hat{\varphi}))$

Figure 11. Helper functions.

8.3 Subsumption and type scheme unification

The subsumption algorithm $subsume(Q, \sigma, \varphi)$ in Figure 10 takes a prefix Q, an F-type σ and a flexible type φ, and returns a new prefix Q' and a substitution θ, such that $Q' \vdash \theta\varphi \sqsubseteq \theta\sigma$. The algorithm is very similar to HMF subsumption (Leijen 2008a). It first skolemizes the σ type and instantiates φ, and unifies the remaining unquantified types. Next, we need to ensure that no skolem constants escape into the environment which is done by checking that the relevant parts of the prefix and substitution do not contain any introduced skolem constant.

Finally, $unifyScheme(Q, \varphi_1, \varphi_2)$ in Figure 10 unifies two flexible types and returns a new prefix Q', a subsitution θ and a type φ such that φ is the most general type where $Q' \vdash \theta\varphi_1 \sqsubseteq \varphi$ and $Q' \vdash \theta\varphi_2 \sqsubseteq \varphi$. Note how the *split* function is used here to generalize over the result type φ.

9. Type inference

The type inference algorithm for HML is defined in Figure 12 and is very similar to MLF inference. It is surprisingly straightforward since the unification algorithm is rather powerful. The rule for variables just looks up the type in the environment. Inference for **let** bindings is very simple and we only need to bind the inferred type for the bound expression in the environment.

For a lambda expression, we assume a fresh type α for the parameter x. After doing inference for the body, we check if the inferred type for x is a monotype. This is to prevent inferring polymorphic types for function parameters as required by the FUN type rule. This check is sufficient because we also ensure that Q only contains polymorphic bounds. Afterwards *split* is used to generalize over the result type. The result type is flexibly quantified

$infer :: (Q, \Gamma, e) \rightarrow (Q, \theta, \varphi)$

$infer(Q, \Gamma, x) =$
  return $(Q,\ [\,],\ \Gamma(x))$

$infer(Q, \Gamma,\ \textbf{let}\ x = e_1\ \textbf{in}\ e_2)$
  let $(Q_1, \theta_1, \varphi_1) = infer(Q, \Gamma, e_1)$
  let $(Q_2, \theta_2, \varphi_2) = infer(Q_1, (\theta_1\Gamma, x : \varphi_1), e_2)$
  return $(Q_2,\ \theta_2 \circ \theta_1,\ \varphi_2)$

$infer(Q, \Gamma, \lambda x.\ e) =$
  assume α, β are fresh
  let $(Q_1, \theta_1, \varphi_1) = infer((Q, \alpha \geqslant \bot), (\Gamma, x : \alpha), e)$
  fail if not $(\theta_1\alpha = \tau)$ for some τ
  let $(Q_2, Q_3) \quad = split(Q_1, dom(Q))$
  let $(Q_3', \theta_3') \quad = extend(Q_3, \beta \geqslant \varphi_1)$
  return $(Q_2,\ \theta_1,\ \forall Q_3'.\ \theta_1\alpha \rightarrow \theta_3'\beta)$

$infer(Q, \Gamma, \lambda(x :: \sigma).\ e) =$
  assume β is fresh, σ is closed
  let $(Q_1, \theta_1, \varphi_1) = infer(Q, (\Gamma, x : \sigma), e)$
  let $(Q_2, Q_3) \quad = split(Q_1, dom(Q))$
  let $(Q_3', \theta_3') \quad = extend(Q_3, \beta \geqslant \varphi_1)$
  return $(Q_2,\ \theta_1,\ \forall Q_3'.\ \sigma \rightarrow \theta_3'\beta)$

$infer(Q, \Gamma, e_1\ e_2) =$
  assume $\alpha_1, \alpha_2, \beta$ are fresh
  let $(Q_1, \varphi_1, \theta_1) = infer(Q, \Gamma, e_1)$
  let $(Q_2, \varphi_2, \theta_2) = infer(Q_1, \theta_1\Gamma, e_2)$
  let $(Q_2', \theta_2') \ = extend(Q_2, \alpha_1 \geqslant \theta_2\varphi_1, \alpha_2 \geqslant \varphi_2, \beta \geqslant \bot)$
  let $(Q_3, \theta_3) \ = unify(Q_2', \theta_2'\alpha_1, \theta_2'\alpha_2 \rightarrow \beta)$
  let $(Q_4, Q_5) = split(Q_3, dom(Q))$
  return $(Q_4,\ \theta_3 \circ \theta_2 \circ \theta_1,\ \forall Q_5.\ \theta_3\beta)$

Figure 12. Type inference.

using $(\beta \geqslant \varphi_1)$. For example, the most general type for the *const* function is:

$$const :: \forall\alpha.\ \forall(\gamma \geqslant \forall\beta.\ \beta \rightarrow \alpha).\ \alpha \rightarrow \gamma \qquad \text{(inferred)}$$
$$const = \lambda x.\ \lambda y.\ x$$

where its type can be instantiate to both of the usual System F types that we can assign to the *const* function:

$$\forall\alpha.\ \forall(\gamma \geqslant \forall\beta.\ \beta \rightarrow \alpha).\ \alpha \rightarrow \gamma \ \sqsubseteq\ \forall\alpha\beta.\ \alpha \rightarrow \beta \rightarrow \alpha$$
$$\forall\alpha.\ \forall(\gamma \geqslant \forall\beta.\ \beta \rightarrow \alpha).\ \alpha \rightarrow \gamma \ \sqsubseteq\ \forall\alpha.\ \alpha \rightarrow (\forall\beta.\ \beta \rightarrow \alpha)$$

This means that we need no special rules for ‘deep instantiation’ or prenex conversions on types (Peyton Jones et al. 2007).

The application case first infers the types φ_1 and φ_2 for the expressions e_1 and e_2. It then unifies the types α_1 and $\alpha_2 \rightarrow \beta$ under a prefix that assumes $(\alpha_1 \geqslant \theta_2\varphi_1)$, $(\alpha_2 \geqslant \varphi_2)$, and an unconstrained $(\beta \geqslant \bot)$. Finally, it generalizes the result using *split*.

Theorem 11 (*Soundness of inference*): If $infer(Q_0, \Gamma, e)$ succeeds with (Q, θ, φ), then $Q, (\theta\Gamma) \vdash \theta e : \theta\varphi$ holds where $Q_0 \sqsubseteq Q$.

Theorem 12 (*Completeness of inference*): Assume a Q_0 where $ftv(\Gamma) \subseteq dom(Q_0)$ and $Q_0 \sqsubseteq Q$. If $Q, \Gamma \vdash e : \varphi$ holds then $infer(Q_0, \Gamma, e)$ succeeds too with a principal solution (Q', θ', φ'), such that $Q' \vdash \theta'\varphi' \sqsubseteq \varphi$ and $Q' \sqsubseteq Q$.

We have done a paper and pencil proof of soundness but we have not done all cases of the completeness proof yet. Since the algorithm is basically a simplification of the usual MLF inference algo-

System	Types	Flexible types on bindings	Annotations
HMF	Regular F-types	No	On polymorphic parameters and ambiguous impredicative applications
FPH	Boxy F-types	No	On polymorphic parameters and some let bindings with higher-rank types
$\mathsf{HML}^{\mathsf{F}}$	Flexible F-types	No	On polymorphic parameters and some let bindings with higher-rank types
HML	Flexible F-types	Let-bound only	On polymorphic parameters
MLF	Flexible and Rigid	Let- and λ-bindings	On parameters that are used polymorphically

Figure 13. A high level comparision between different inference systems.

rithm, the proofs are very similar in structure to the ones for MLF (Le Botlan 2004). One way to potentially simplify the proofs is to formulate the inference algorithm in terms of recent work on MLF unification using a graphical representation of MLF types (Rémy and Yakobowski 2007) but this is still work in progress.

9.1 Implementing $\mathsf{HML}^{\mathsf{F}}$

The $\mathsf{HML}^{\mathsf{F}}$ variant is very straightforward to implement. Since the algorithm by construction always derives most-general types, we just need to call the *ftype* function when binding a let-bound value in the environment:

$$
\begin{array}{l}
\mathit{infer}(Q, \Gamma,\ \mathbf{let}\ x = e_1\ \mathbf{in}\ e_2) \\
\quad \text{let } (Q_1, \theta_1, \varphi_1) = \mathit{infer}(Q, \Gamma, e_1) \\
\quad \text{let } (Q_2, \theta_2, \varphi_2) = \mathit{infer}(Q_1, (\theta_1\Gamma, x : \mathit{ftype}(\varphi_1)), e_2) \\
\quad \text{return } (Q_2,\ \theta_2 \circ \theta_1,\ \varphi_2)
\end{array}
$$

and no further changes are required.

10. Related work

MLF was first described by by Rémy and Le Botlan (2004; 2003; 2007; 2007). The extension of MLF with qualified types is described in (Leijen and Löh 2005). Leijen later gives a type directed translation of MLF to System F and describes Rigid-MLF (Leijen 2007), a variant of MLF that does not assign polymorphically bounded types to let-bound values but internally still needs the full inference algorithm of MLF.

Rémy and Le Botlan introduce Implicit MLF (2007) as a simpler version of MLF where only let-bindings can have types with flexible quantification. This simplification allows them to give a semantics of flexible types in terms of sets of System F types. Unfortunately, inference for Implicit MLF is undecidable. The variant with type annotations is introduced as *e*MLF and uses rigid quantification to enable type inference.

Leijen describes HMF (Leijen 2008a): a type inference system for first-class polymorphism that is based on regular Hindley-Milner inference. Annotations are required on function parameters with a polymorphic type and on all ambiguous impredicative applications. Even though the annotation rule is harder than that of HML or MLF, it represents an interesting design point since the specification and the type inference algorithm are very simple: the declarative type rules use only regular System F types and the inference algorithm is a small extension of algorithm W.

Vytiniotis et al. (2006) describe boxy type inference which is made principal by distinguishing between inferred 'boxy' types, and checked annotated types. A critique of boxy type inference is that its specification has a strong algorithmic flavor which can make it fragile under small program transformations (Rémy 2005).

The FPH type system (Vytiniotis et al. 2008) is an improvement on earlier work and also uses boxy types to enable type inference for first-class polymorphism. FPH only requires annotations on polymorphic parameters and may require an annotation for let-bindings with a higher-rank type. See also Section 7.1 for a more detailed comparison with HML.

As a reference, Figure 13 gives a high-level comparison between these proposed type systems for first-class polymorphism. Note that type inference for HMF is significantly simpler than for the other systems in this table. Also, systems that allow flexible types on let-bindings (MLF and HML) add true expressiveness since they allow the sharing of values at different polymorphic instances (as shown in Section 2.1 and Section 7).

To the best of our knowledge, a type inference algorithm for the simply typed lambda calculus was first described by Curry and Feys (1958). Later, Hindley (1969) introduced the notion of principal type, proving that the Curry and Feys algorithm inferred most general types. Milner (1978) independently described a similar algorithm, but also introduced the important notion of first-order polymorphism where let-bound values can have a polymorphic type. Damas and Milner (1982) later proved the completeness of Milner's algorithm, extending the type inference system with polymorphic references (Damas 1985). Wells (1999) shows that general type inference for unannotated System F is undecidable.

Jones (1997) extends Hindley-Milner with first class polymorphism by wrapping polymorphic values into type constructors. This is a simple and effective technique that is widely used in Haskell but one needs to define a special constructor and operations for every polymorphic type. Garrigue and Rémy (1999) use a similar technique but can use a generic 'box' operation to wrap polymorphic types. Odersky and Läufer (1996) describe a type system that has higher-rank types but no impredicative instantiation. Peyton Jones et al. (2007) extend this work with type annotation propagation. Dijkstra (2005) extends this further with bidirectional annotation propagation to support impredicative instantiation.

11. Conclusion

We presented HML, a type inference system that supports full first-class polymorphism with few annotations: only function parameters with a polymorphic type need to be annotated. It has logical type rules with principal derivations where flexible types ensure that every expression can be assigned a most general type. The type inference algorithm is straightforward and we have implemented a reference implementation that represents substitutions using updateable references and uses ranked type variables to do efficient generalization (Leijen 2008b).

Given these good properties of HML, we think it can be an excellent type system in practice for languages that support first-class polymorphic values.

References

H. Curry and R. Feys. *Combinatory Logic*, volume 1. North-Holland, 1958.

Luis Damas. *Type Assignment in Programming Languages*. PhD thesis, University of Edinburgh, April 1985. Technical report CST-33-85.

Luis Damas and Robin Milner. Principal type-schemes for functional programs. In *9th ACM symp. on Principles of Program-*

ming Languages (POPL'82), pages 207–212, 1982.

Atze Dijkstra. *Stepping through Haskell*. PhD thesis, Universiteit Utrecht, Nov. 2005.

Jacques Garrigue and Didier Rémy. Semi-explicit first-class polymorphism for ML. *Journal of Information and Computation*, 155:134–169, 1999.

J.R. Hindley. The principal type scheme of an object in combinatory logic. *Transactions of the American Mathematical Society*, 146:29–60, Dec. 1969.

Mark P. Jones. First-class polymorphism with type inference. In *24th ACM Symposium on Principles of Programming Languages (POPL'97)*, January 1997.

George Kuan and David MacQueen. Efficient ML type inference using ranked type variables. In *The 2007 ACM SIGPLAN Workshop on ML (ML 2007)*, Freiburg, Germany, October 2007.

Didier Le Botlan. *ML^F: Une extension de ML avec polymorphisme de second ordre et instanciation implicite*. PhD thesis, INRIA Rocquencourt, May 2004. Also in English.

Didier Le Botlan and Didier Rémy. MLF: Raising ML to the power of System-F. In *The International Conference on Functional Programming (ICFP'03)*, pages 27–38, aug 2003.

Didier Le Botlan and Didier Rémy. Recasting MLF. Research Report 6228, INRIA, Rocquencourt, BP 105, 78 153 Le Chesnay Cedex, France, June 2007.

Daan Leijen. HMF: Simple type inference for first-class polymorphism. In *13th ACM symp. of the International Conference on Functional Programming (ICFP'08)*, September 2008a. Extended version available as a technical report: MSR-TR-2007-118, Sep 2007, Microsoft Research.

Daan Leijen. A reference implementation of HML. Available at `http://research.microsoft.com/users/daan/pubs.html`, April 2008b.

Daan Leijen. A type directed translation from MLF to System F. In *The International Conference on Functional Programming (ICFP'07)*, Oct. 2007.

Daan Leijen and Andres Löh. Qualified types for MLF. In *The International Conference on Functional Programming (ICFP'05)*. ACM Press, Sep. 2005.

Robin Milner. A theory of type polymorphism in programming. *Journal of Computer and System Sciences*, 17:248–375, 1978.

Martin Odersky and Konstantin Läufer. Putting type annotations to work. In *23th ACM symp. on Principles of Programming Languages (POPL'96)*, pages 54–67, January 1996.

Simon Peyton Jones, Dimitrios Vytiniotis, Stephanie Weirich, and Mark Shields. Practical type inference for arbitrary-rank types. *Journal of Functional Programming*, 17(1):1–82, 2007.

Didier Rémy. Simple, partial type-inference for System-F based on type-containment. In *The International Conference on Functional Programming (ICFP'05)*, September 2005.

Didier Rémy and Boris Yakobowski. A graphical presentation of MLF types with a linear-time unification algorithm. In *TLDI'07*, pages 27–38, 2007.

Dimitrios Vytiniotis, Stephanie Weirich, and Simon Peyton Jones. Boxy types: type inference for higher-rank types and impredicativity. In *The International Conference on Functional Programming (ICFP'06)*, September 2006.

Dimitrios Vytiniotis, Stephanie Weirich, and Simon Peyton Jones. FPH : First-class polymorphism for Haskell. In *13th ACM symp. of the International Conference on Functional Programming (ICFP'08)*, September 2008.

J.B. Wells. Typability and type checking in System-F are equivalent and undecidable. *Ann. Pure Appl. Logic*, 98(1–3):111–156, 1999.

$$\text{VAR-G} \quad \frac{x : \varphi \in \Gamma}{\Gamma \vdash x : \varphi}$$

$$\text{INST-G} \quad \frac{\Gamma \vdash e : \varphi_1 \quad \varphi_1 \sqsubseteq \varphi_2}{\Gamma \vdash e : \varphi_2}$$

$$\text{GEN-G} \quad \frac{\Gamma \vdash e : \varphi_2 \quad \alpha \notin \mathit{ftv}(\Gamma)}{\Gamma \vdash e : \forall\alpha.\, \varphi_2}$$

$$\text{APP-G} \quad \frac{\Gamma \vdash e_1 : \forall Q.\, \sigma_2 \to \sigma \quad \Gamma \vdash e_2 : \forall Q.\, \sigma_2}{\Gamma \vdash e_1\ e_2 : \forall Q.\, \sigma}$$

$$\text{LET-G} \quad \frac{\Gamma \vdash e_1 : \varphi_1 \quad (\Gamma, x : \varphi_1) \vdash e_2 : \varphi_2}{\Gamma \vdash \textbf{ let } x = e_1 \textbf{ in } e_2 : \varphi_2}$$

$$\text{FUN-G} \quad \frac{(\Gamma, x : \tau) \vdash e : \varphi \quad \alpha \notin \mathit{ftv}(\tau)}{\Gamma \vdash \lambda x.\, e : \forall(\alpha \geqslant \varphi).\, \tau \to \alpha}$$

$$\text{FUN-ANN-G} \quad \frac{(\Gamma, x : \sigma) \vdash e : \varphi \quad \alpha \notin \mathit{ftv}(\sigma)}{\Gamma \vdash \lambda(x :: \sigma).\, e : \forall(\alpha \geqslant \varphi).\, \sigma \to \alpha}$$

Figure 14. Type rules for HML^{G}.

A. Restricting parameter types

As shown in Section 4, the principal type of the expression $\lambda x.\, \mathit{choose}\ \mathit{id}\ x$ is $\forall(\beta \geqslant \forall\alpha.\, \alpha \to \alpha).\, \beta \to \beta$. This type can be instantiated to either the Hindley-Milner type $\forall\alpha.\, (\alpha \to \alpha) \to \alpha \to \alpha$, or the System F type $(\forall\alpha.\, \alpha \to \alpha) \to (\forall\alpha.\, \alpha \to \alpha)$. The latter type is somewhat surprising – are we able to infer a polymorphic type for the parameter after all? Of course, the derivation is sound since we are unable to *use* the parameter polymorphically in the body of the lambda expression. In that sense, the instantiation is equivalent to any other impredicative instantiation.

Nevertheless, one may want to disallow derivations where the free type variables in a parameter type have a polymorphic bound. We call this system HML^{G}. The only change to the type rules of HML would be in the FUN rule:

$$\text{FUN-MONO} \quad \frac{Q, (\Gamma, x : \tau) \vdash e : \sigma \quad \forall Q.\, \tau \equiv \forall\overline{\alpha}.\, \tau}{Q, \Gamma \vdash \lambda x.\, e : \tau \to \sigma}$$

With this rule, the principal type of $\lambda x.\, \mathit{choose}\ \mathit{id}\ x$ becomes the Hindley-Milner type $\forall\alpha.\, (\alpha \to \alpha) \to \alpha \to \alpha$.

One may feel that this restriction is not very useful in practice but it has an interesting side effect that with this restriction there is no need anymore to state the type rules and type instantiation under a prefix! Figure 14 defines HML^{G} formally where we no longer need an explicit prefix. The rules VAR-G, INST-G, and GEN-G stay the same except that instantiation is no longer stated under a prefix. The application rule APP-G now requires that the types of the function and argument have the same quantifiers Q. This rule is also applicable in HML (and can be derived directly (Le Botlan 2004)). The rule for let bindings is unchanged, and the rules for functions simply generalize directly over the result type.

Of course, even though the HML^{G} is simpler than HML, it is also less general. In particular, an η-expanded expression may have a less general types than the original expression which may be undesirable in practice.

B. Back to System F

Interestingly, we can combine HML^{F} and HML^{G} such that we don't even need flexible types anymore. Figure 15 describes the type rules of HML^{FG}. The type rules should look very familiar since they are just the rules of System F, extended with a rule for let bindings, annotated functions, and annotated let bindings. Even though this is a restriction of HML, it may be interesting in practice as programmers are only exposed to regular System F types and can reason about type derivations using just System F types!

$$\text{VAR-FG} \quad \frac{x : \sigma \in \Gamma}{\Gamma \vdash x : \sigma}$$

$$\text{INST-FG} \quad \frac{\Gamma \vdash e : \sigma_1 \quad \sigma_1 \sqsubseteq_{\mathsf{F}} \sigma_2}{\Gamma \vdash e : \sigma_2}$$

$$\text{GEN-FG} \quad \frac{\Gamma \vdash e : \sigma \quad \alpha \notin \mathit{ftv}(\Gamma)}{\Gamma \vdash e : \forall\alpha.\, \sigma}$$

$$\text{APP-FG} \quad \frac{\Gamma \vdash e_1 : \sigma_2 \to \sigma \quad \Gamma \vdash e_2 : \sigma_2}{\Gamma \vdash e_1\ e_2 : \sigma}$$

$$\text{FUN-FG} \quad \frac{\Gamma, x : \tau \vdash e : \sigma}{\Gamma \vdash \lambda x.\, e : \tau \to \sigma}$$

$$\text{LET-FG} \quad \frac{\begin{array}{c}\Gamma \vdash e_1 : \sigma_1 \quad \mathit{minimal}(\sigma_1) \\ \Gamma, x : \sigma_1 \vdash e_2 : \sigma_2\end{array}}{\Gamma \vdash \textbf{ let } x = e_1 \textbf{ in } e_2 : \sigma_2}$$

$$\text{LET-ANN-FG} \quad \frac{\begin{array}{c}\Gamma \vdash e_1 : \sigma_1 \\ \Gamma, x : \sigma_1 \vdash e_2 : \sigma_2\end{array}}{\Gamma \vdash \textbf{ let } x : \sigma_1 = e_1 \textbf{ in } e_2 : \sigma_2}$$

$$\text{FUN-ANN-FG} \quad \frac{\Gamma, x : \sigma_1 \vdash e : \sigma_2}{\Gamma \vdash \lambda(x :: \sigma_1).\, e : \sigma_1 \to \sigma_2}$$

Figure 15. Type rules for HML^{FG}.

Of course, by giving up flexible types, we do not have principal type derivations anymore. For example, the expression $\mathit{single}\ \mathit{id}$ can have both the type $\mathit{List}\ (\forall\alpha.\, \alpha \to \alpha)$ and $\forall\alpha.\, \mathit{List}\ (\alpha \to \alpha)$.

However, we are going to ensure that any let-bound expression has a principal type through the rule LET-FG. In particular, the constraint $\mathit{minimal}(\sigma_1)$ should ensure the we can infer a unique principal type for the let-bound expression. We have some choice here for the definition of this constraint. A conservative choice would be to ensure that σ_1 is a Hindley-Milner rank-1 type, i.e.

$$\mathit{minimal}(\sigma) \;\Leftrightarrow\; \sigma \equiv \forall\overline{\alpha}.\, \tau$$

By choosing this definition, let-bound expressions with a higher rank type must be annotated (together with parameters with a polymorphic type). A similar type system as this has been described too as a restriction of boxy types (Vytiniotis et al. 2008). The inference algorithm of HML (Section 9) can readily be adapted to use the *ftype* function as described in Section 9.1 to instantiate the type of a let binding, and simply reject let-bound expressions with a higher-rank type. Note that even though the type rules just use System F types, the type inference algorithm still uses flexible types internally to represent the set of derivable System F types efficiently.

Unfortunately, the above restriction rejects many programs with unique higher-rank types, for example:

$\textbf{let}\ \mathit{xs} = \mathit{cons}\ \mathit{id}\ \mathit{ids}$

We can also formulate a minimality constraint as in HML^{F} where we only allow the most general System F type with the least number of inner quantifiers (or lowest rank). This variant can be implemented very easily by just applying the *ftype* function on the inferred let-bound expression as described in Section 9.1. We leave a proper investigation of these different variants to future work.