Moscow ML Owner's Manual

Sergei Romanenko and Peter Sestoft

Version 1.31 of 15 October 1995

Moscow ML implements a subset of the Standard ML Core language as defined in the *Definition of Standard ML* [6], except that a few built-in functions have changed to fit the new SML Standard Library.

Moscow ML also provides a simple subset of the Standard ML Modules language, restricted to signatures and non-nested structures. It supports separate compilation and the generation of stand-alone executables in a straightforward manner.

This document explains how to use the Moscow ML system, and describes the built-in functions available at top-level. The Standard ML language itself is not documented here. Section 11 lists textbooks and other material that describe the language. In addition, a compact summary of Moscow ML syntax and built-in functions can be found in the *Moscow ML Language Overview* [7].

Acknowledgements: The Caml Light system was instrumental in creating Moscow ML, which uses the Caml Light runtime system and essentially the same bytecode generator. Many other aspects of the design were also derived from Caml Light, developed by Xavier Leroy and Damien Doligez at INRIA Rocquencourt, France [3, 4].

The ML Kit helped solving problems of parsing, infix resolution, and type inference [2].

The Moscow ML home page is at `http://www.dina.kvl.dk/~sestoft/mosml.html`.

Contents

1 Getting started

1.1 Installation

Get a copy of the Moscow ML system executables (see Section 10 for instructions) and unpack them in your home directory (under Unix) or in directory `C:\` (under DOS). This creates a directory `mosml`. Read the file `mosml/install.txt`. This manual, and the *Moscow ML Language Overview*, are in directory `mosml/doc`.

1.2 The interactive system

The interactive system is invoked by typing `mosml` at the shell prompt. It allows you to enter declarations and evaluate expressions:

```
> mosml
Moscow ML version 1.31 (15 October 1995)
Enter `quit();´ to quit.
- fun fac n = if n = 0 then 1 else n * fac (n-1);
> val fac = fn : int -> int
- fac 10;
> val it = 3628800 : int
-
```

You can quit the interactive session by typing '`quit();`' or control-D (under Unix) or control-Z followed by newline (under DOS). Type `help "lib";` for an overview of built-in function libraries, and e.g. `help "Array"` for help on Array operations. See Section 3 for further information on `mosml`.

1.3 New SML Standard Library

Moscow ML version 1.31 implements a large part of the SML Standard Library, which is intended to enhance the portability of Standard ML programs[1]. Future releases of Standard ML of New Jersey, and of Moscow ML, will implement the SML Standard Library. Since the SML Standard Library is still under development, the current implementation must be considered non-authoritative. In particular, the new Standard Library input/output functions are not yet implemented in Moscow ML.

The new SML Standard Library is slightly incompatible with both the *Definition of Standard ML* and with SML/NJ version 0.93. Invoking Moscow ML with '`mosml -P oldbasis`' gives the old versions of `chr`, `explode`, `implode`, and `ord`, which work on strings. Invoking Moscow ML with option '`mosml -P nj93`', gives an initial environment more compatible with the old SML/NJ version 0.93. See Section 3.1 for a detailed explanation of the `-P` option.

1.4 The batch compiler and linker

The batch compiler and linker is invoked by typing `mosmlc` at the shell prompt. It can compile ML source files separately (`mosmlc -c`) and link them to obtain executables (`mosmlc -o`), in a manner similar to C compilers. The compatibility options `-P oldbasis` and `-P nj93` mentioned above can be used with the batch compiler also. See Section 5 for further information on `mosmlc`.

1.5 A simple module system

Moscow ML also provides a simple subset of the Standard ML Modules language, restricted to signatures and non-nested structures. A Moscow ML program consists of one or more units. A unit U has a signature (or interface) in file U.sig and a body (or implementation) in file U.sml. The unit signature corresponds

[1]The SML Standard Library is being developed by Andrew Appel (Princeton University, New Jersey, USA); Emden Gansner, Lal George, Lorenz Huelsbergen, Dave MacQueen, John Reppy (AT&T Bell Labs, New Jersey, USA); Matthew Arcus, Dave Berry, Richard Brooksby, Nick Barnes, Brian Monahan, Jon Thackray (Harlequin Ltd., Cambridge, England); Carsten Müller (Berlin); and Peter Sestoft (Royal Veterinary and Agricultural University, Denmark).

to a Standard ML signature, and the unit body corresponds to a Standard ML structure. Moscow ML supports type-safe separate compilation and linking.

Section 4 gives the syntax and an informal semantics of compilation units. Section 6 explains automatic recompilation management.

1.6 What is new in release 1.31

- Installable prettyprinters are provided.
- Some minor bugs and deficiencies have been fixed; see file `doc/releases.txt`.

1.7 What was new in release 1.30

- The system has been rewritten in Standard ML and can compile itself. This makes installation under Unix much easier than it used to be.
- The garbage collector of the Caml Light runtime system has been improved (by Damien Doligez at INRIA).
- The new initial basis for Standard ML has been adopted. A new primitive type `char` has been introduced, the types of `chr`, `explode`, `implode` and `ord` have changed, and several new utility libraries have been added.
- Quotations and antiquotations are supported.

2 The initial environment

Moscow ML does not yet fully adhere to the new Standard Library. In particular, the new input-output operations have not been implemented yet.

2.1 Built-in variables and functions

For each variable we list its type, effect, and any exception it may raise.

An identifier with type involving `num` stands for two functions: one in which `num` (in the type) is replaced by `int`, and another in which `num` is replaced by `real`. Similarly for `cirs`, which stands for `char`, `int`, `real`, or `string`. An explicit type constraint will be needed if the surrounding text does not determine the type of a particular occurrence.

Nonfix identifiers in the initial environment

id	type	effect	exception
~	num -> num	arithmetic negation	Overflow
!	´a ref -> ´a	dereference	
abs	num -> num	absolute value	Overflow
app	(´a -> unit) -> ´a list -> unit	apply to all elements	
arctan	real -> real	arc tangent (to radians)	
chr	int -> char	character with number	Chr
concat	string list -> string	concatenate strings	Size
cos	real -> real	cosine (from radians)	Fail
exp	real -> real	exponential function	Overflow
explode	string -> char list	list of characters in string	
false	bool	logical falsehood	
floor	real -> int	round towards $-\infty$	Overflow
foldl	(´a*´b->´b)->´b->´a list->´b	fold from left to right	
foldr	(´a*´b->´b)->´b->´a list->´b	fold from right to left	
hd	´a list -> ´a	first element	Empty
ignore	´a -> unit	discard argument	
implode	char list -> string	make string from characters	Size
length	´a list -> int	length of list	
map	(´a -> ´b) -> ´a list -> ´b list	map over all elements	
makestring	cirs -> string	convert to string	
ln	real -> real	natural logarithm	Ln
nil	´a list	empty list	
not	bool -> bool	logical negation	
null	´a list -> bool	true if list is empty	
ord	char -> int	number of character	
real	int -> real	int to real	
ref	´_a -> ´_a ref	create reference value	
rev	´a list -> ´a list	reverse list	
sin	real -> real	sine (from radians)	Fail
size	string -> int	length of string	
sqrt	real -> real	square root	Sqrt
str	char -> string	create one-character string	
tl	´a list -> ´a list	tail of list	Empty
true	bool	logical truth	

Infixed identifiers in the initial environment

id	type	effect	exception
Infix precedence 7:			
`/`	`real * real -> real`	floating-point quotient	`Div, Overflow`
`div`	`int * int -> int`	quotient (round towards $-\infty$)	`Div, Overflow`
`mod`	`int * int -> int`	remainder (of `div`)	`Div, Overflow`
`quot`	`int * int -> int`	quotient (round towards 0)	`Div, Overflow`
`rem`	`int * int -> int`	remainder (of `quot`)	`Div, Overflow`
`*`	`num * num -> num`	product	`Overflow`
Infix precedence 6:			
`+`	`num * num -> num`	sum	`Overflow`
`-`	`num * num -> num`	difference	`Overflow`
`^`	`string * string -> string`	concatenate	`Size`
Infix precedence 5:			
`::`	`'a * 'a list -> 'a list`	cons onto list (R)	
`@`	`'a list * 'a list -> 'a list`	append lists (R)	
Infix precedence 4:			
`=`	`''a * ''a -> bool`	equal to	
`<>`	`''a * ''a -> bool`	not equal to	
`<`	`cirs * cirs -> bool`	less than	
`<=`	`cirs * cirs -> bool`	less than or equal to	
`>`	`cirs * cirs -> bool`	greater than	
`>=`	`cirs * cirs -> bool`	greater than or equal to	
Infix precedence 3:			
`:=`	`'a ref * 'a -> unit`	assignment	
`o`	`('b->'c) * ('a->'b) -> ('a->'c)`	function composition	
Infix precedence 0:			
`before`	`'a * unit -> 'a`	discard second argument	

2.2 Non-standard types and literals

In keeping with the new Standard Library, we provide support for the new type constructors `char` and `vector`.

`char`

`char` is the type of characters. A character constant consists of the symbol # immediately followed by a string constant of length one. Examples are `#"a"`, `#"\n"`, `#"\^Z"`, `#"\255"`, and `#"\""`. Non-examples are `# "a"` and `#c`. Operations on characters are defined in unit `Char`.

`vector`

`'a vector` is the type of immutable vectors with elements of type `'a`. A vector expression is a list expression immediately preceded by the symbol '`#`', that is `#[`exp_1`,`$\cdots$`,`exp_n`]`. Similarly for vector patterns `#[`pat_1`,`$\cdots$`,`pat_n`]`. Operations on vectors are defined in unit `Vector`.

2.3 General non-standard primitives

The following general non-standard primitives, and most of the standard ones from the Definition, are defined in unit `General`, which is open by default.

`makestring : 'a -> string`

`makestring` is an overloaded function whose argument must be resolved as either `char`, `int`, `real`, or `string`. It returns the string that `print` would produce on the screen.

`op <, >, <=, >=`

The overloaded operators `<`, `>`, `<=`, and `>=` can be used to compare characters and strings lexicographically (in addition to integers and reals).

`exception Fail of string`
: Raised by `compile` and `load` to report errors; see Section 3.2. Also raised by `sin` and `cos` if the argument is so large that no meaningful result can be returned.

`exception Size and Subscript and Overflow`
: Raised by primitives in e.g. the `Array`, `List`, and `Vector` standard libraries.

`exception Graphic_failure of string` Raises if error occurs during execution of a graphics primitive (DOS version only).

`datatype 'a option = NONE | SOME of 'a`
`datatype ordering = LESS | EQUAL | GREATER`
: Some useful datatypes for partial functions and string comparisons.

`datatype 'a frag = QUOTE of string | ANTIQUOTE of 'a`
: Used to represent the value of a quotation; see Section 2.6.

2.4 Non-standard input-output primitives

The following non-standard input-output primitives, and the standard ones from the Definition, are defined in unit `BasicIO`, which is open by default. (Unit `BasicIO` will eventually be replaced by units `TextIO` and `BinIO`, as required by the new Standard Library).

`exit : int -> 'a`
: Evaluating `exit(`n`)` quits Moscow ML with exit code n. Use $n = 0$ for normal termination, and $n > 0$ for abnormal termination.

`flush_out : outstream -> unit`
: Evaluating `flush_out(`os`)` ensures that the consumer associated with outstream os has received all of the characters that have been written to os.

`input_line : instream -> string`
: Evaluating `input_line(`is`)` returns a string consisting of the characters from is up through, and including, the next end-of-line character. If the end of stream is reached without reaching an end-of-line character, all remaining characters from the stream (without an end-of-line character) are returned.

`inputc : instream -> int -> string`
: Evaluating `inputc` is n has the same effect as `input(`is`,`n`)`.

`open_append : string -> outstream`
: A file named f may be opened for output while preserving its previous contents by opening it with `open_append("`f`")`. Output to the outstream returned by `open_append` is appended to the previous contents of the file.

`outputc : outstream -> string -> unit`
: Evaluating `outputc` os s has the same effect as `output(`os`,`s`)`.

`say : string -> unit`
: Evaluating `say` s has the same effect as `(output(std_out,`s`); flush_out std_out)`.

`std_err : outstream`
: This stream corresponds to 'standard error' in C. All output to `std_err` is flushed immediately.

2.5 Library structures

Unit name	Description	Status
`Array`	Mutable polymorphic arrays	SDF
`Bool`	Booleans	S F
`Byte`	Conversions between `Word8` and `Char`	S F
`Char`	Characters	SDF
`CharArray`	Mutable arrays of characters	S F
`CharVector`	Immutable vectors of characters (that is, strings)	S F
`Date`	From time points to dates and vice versa	S F
`FileSys`	File system interface	S F
`Graphics`	Graphics primitives (DOS only)	
`Integer`	Integer arithmetic and comparisons	S F
`List`	Lists	SDF
`ListPair`	Pairs of lists	S F
`Math`	Trigonometric and transcendental functions	S F
`OS`	Operating system interface	S F
`Path`	File pathnames	S F
`PP`	General prettyprinters	F
`Process`	Process interface	S F
`Real`	Real arithmetic and comparisons	S F
`String`	String utilities	SDF
`StringCvt`	Conversion to and from strings	S F
`Substring`	Scanning of substrings	S F
`Time`	Time points and durations	S F
`Timer`	Timing operations	S F
`Vector`	Immutable vectors	SDF
`Word`	Unsigned 31-bit integers ('machine words')	S F
`Word8`	Unsigned 8-bit integers (bytes)	S F
`Word8Array`	Mutable arrays of unsigned 8-bit integers	S F
`Word8Vector`	Immutable vectors of unsigned 8-bit integers	S F

Only the libraries marked `S` belong to the SML Standard Library; the remaining ones are non-standard.

`S` means that the unit belongs to the SML Standard Library.
`D` means that the unit is loaded by default.
`F` means that the unit is loaded when option `-P full` is specified.

The *Moscow ML Language Overview* lists the functions defined by each of these structures. The structures are thoroughly described in the Standard Library document [1], and in the signature files in directory `mosml/lib`. For instance, the `List` functions are described in file `mosml/lib/List.sig`. A simple on-line library browser can be invoked by typing e.g. `help "List";`.

2.6 Quotations and antiquotations

As in Standard ML of New Jersey, so-called *quotations* can be used to embed object language phrases in ML programs. Quotations are disabled by default. To enable quotations in the interactive system (`mosml`), execute `quotation := true`. This allows quotations to appear in declarations entered at top-level and in files compiled by the primitive `compile`. To enable quotations in files compiled with the batch compiler `mosmlc`, invoke it with option `-q` as in `mosmlc -q`.

A quotation is a particular kind of expression and consists of a non-empty sequence of (possibly empty) *fragments* surrounded by backquotes:

exp	::=	`` `frags` ``	quotation
frags	::=	*charseq*	character sequence
		charseq `^`*id frags*	antiquotation variable
		charseq `^(`*exp*`)` *frags*	antiquotation expression

The *charseq* is a possibly empty sequence of printable characters or spaces or tabs or newlines. A quotation evaluates to a value of type `ty frag list` where `ty` is the type of the antiquotation variables and antiquotation expressions. The type `'a frag` is from unit `General`; see Section 2.3.

A *charseq* fragment evaluates to `QUOTE "`*charseq*`"`. An antiquotation fragment `^`*id* or `^(`*exp*`)` evaluates to `ANTIQUOTE` *value* where *value* is the value of the variable *id* resp. the expression *exp*. All antiquotations in a quotation must have the same type `ty`.

An antiquotation fragment is always surrounded by (possibly empty) quotation fragments; and no two quotation fragments can be adjacent. The entire quotation is parsed before any antiquotation inside it is evaluated. Hence changing the value of `Meta.quotation` in an antiquotation inside a quotation has no effect on the parsing of the containing quotation.

For an example, say we have written an ML program to analyse C program phrases, and that we want to enter the C declaration `char s[6] = "abcde"`. We could simply define it as a string:

```
val phrase = "char s[6] = \"abcde\"";
```

but then we need to escape the quotes (") in the C declaration, which is tiresome. If instead we use a quotation, these escapes are not needed:

```
val phrase = `char s[6] = "abcde"`;
```

It evaluates to `[QUOTE "char s[6] = \"abcde\""] : 'a frag list`. Moreover, suppose we want to generate such declarations for other strings than just `"abcde"`, and that we have an abstract syntax for C phrases:

```
datatype cprog =
    IntCst of int
  | StrCst of string;
  | ...
```

Then we may replace the string `"abcde"` by an antiquotation `^(StrCst str)`, and the array dimension 6 by an antiquotation `^(IntCst (size str + 1))`, and make the string `str` a function parameter:

```
fun mkphrase str = `char s[^(IntCst (size str + 1))] = ^(StrCst str)`;
```

Evaluating `mkphrase "longer"` produces the following representation of a C phrase:

```
[QUOTE "char s[", ANTIQUOTE (IntCst 7), QUOTE "] = ",
 ANTIQUOTE (StrCst "longer"), QUOTE ""] : cprog frag list
```

3 The interactive system

The interactive system `mosml` is invoked simply by typing `mosml` at the command line:

```
> mosml
Moscow ML version 1.31 (15 October 1995)
Enter `quit();' to quit.
-
```

The interactive system can be terminated by typing `quit();` and newline, or control-D (under Unix) or control-Z and newline (under DOS). Type '`help "";`' for help on built-in functions.

Invoking the interactive system with command line arguments

`mosml` file$_1$ ... file$_n$

is equivalent to invoking `mosml` and, when Moscow ML has started, entering

```
use "file₁"; ...; use "fileₙ";
```

3.1 Command-line options

`-I` directory
: Specify the default directories to be searched by `use`, `compile` and `load`. A call to `use`, `compile` or `load` will first search the current directory, then all directories specified by option '`-I`' in order of appearance from left to right, and finally the standard library directory.

`-P` unit-set
: Determines which library units will be included and open at compile-time, and which library units will be linked in at link-time:

  - `-P default` This provides an initial environment for the new Standard Library, compatible with future versions of SML/NJ. The units `Array`, `Char`, `List`, `String`, and `Vector` will be loaded, and units `Char`, `List`, and `String` will be partially opened, see Section 2.1. This is the default.
  - `-P oldbasis` This provides an initial environment which is upwards compatible with that of the *Definition of Standard ML* and with pre-1.30 releases of Moscow ML. In particular, the functions `chr`, `explode`, `implode`, and `ord` work on strings, not characters. The new versions of these functions are still available as `Char.chr`, `Char.ord`, `String.explode`, and `String.implode`. In addition the same (new) libraries are loaded as with `-P default`.
  - `-P nj93` This provides an top-level environment which is mostly compatible with that of SML/NJ 0.93. The functions `app`, `ceiling`, `chr`, `dec`, `explode`, `fold`, `hd`, `implode`, `inc`, `max`, `min`, `nth`, `nthtail`, `ord`, `ordof`, `revapp`, `revfold`, `substring`, `tl`, and `truncate` have the same type and meaning as in SML/NJ 0.93. Note that this is incompatible with future versions of SML/NJ. In addition the same (new) libraries are loaded as with `-P default`.
  - `-P full` This loads all the libraries marked `F` in Section 2.5, and partially opens the `Char`, `List`, and `String` units (see Section 2.1).
  - `-P none` No library units are loaded or opened initially.

  Additional library units can loaded into the interactive system by using the `load` function; see Section 3.2 below.

`-stdlib` stdlib-directory
: Specify the standard library directory to be stdlib-directory. The default standard library is usually `mosml/lib` under Unix and `c:\mosml\lib` under DOS.

3.2 Non-standard primitives in the interactive system

The following non-standard primitives are defined in unit `Meta`, loaded (and open by default) only in the interactive system. Hence these primitives cannot be used from source files which are compiled separately. The functions `compile` and `load` deal with Moscow ML compilation units; see Section 4.

`print : 'a -> 'a`
: `print` is a polymorphic function provided as a quick debugging aid. This is an identity function, which as a side-effect prints its argument to standard output exactly as it would be printed at top-level. Output is flushed immediately. For printing simple strings, the function `say` is probably more useful than `print`; see below.

`printDepth : int ref`
: This variable determines the depth (in terms of nested constructors, records, tuples, lists, and vectors) to which values are printed by the top-level value printer and the function `print`. The components of the value whose depth is greater than `printDepth` are printed as '`#`'. The initial value of `printDepth` is 20. This value can be changed by the user's program at any moment.

`printLength : int ref`
: This variable determines the way in which list values are printed by the top-level value printer and the function `print`. If the length of a list is greater than `printLength`, only the first `printLength` elements are printed, and the remaining elements are printed as '...'. The initial value of `printLength` is 200. This value can be changed by the user's program at any moment.

`quit : unit -> 'a`
: Evaluating `quit()` quits Moscow ML immediately.

`installPP : (ppstream -> 'a -> unit) -> unit`
: Evaluating `installPP` *pp* installs the prettyprinter *pp* at type `ty`, provided *pp* has type `ppstream -> ty -> unit`. The type `ty` must be a nullary (parameter-less) type constructor representing a datatype, either built-in (such as bool) or user-defined. Whenever a value of type `ty` is about to be printed by the interactive system, and whenever function `print` is invoked on an argument of type `ty`, the prettyprinter *pp* will be invoked to print it.
: The prettyprinter *pp* must not call the `flush_ppstream` function. See library unit `PP` for more information on prettyprinters.

`use : string -> unit`
: Evaluating `use "`*f*`"` causes ML declarations to be read from file *f* as if they were entered from the console. A file loaded by `use` may, in turn, evaluate calls to `use`. For best results, use `use` only at top level, or at top level within a `use`'d file.

`compile : string -> unit`
: Evaluating `compile "U.sig"` will compile and elaborate the unit signature in file U.sig, producing a compiled signature file U.ui. During compilation, the compiled signatures of other units will be accessed if they are mentioned in U.sig.
: Evaluating `compile "U.sml"` will elaborate and compile the unit body in file U.sml, producing a bytecode file U.uo. If there is an explicit signature U.sig, then file U.ui must exist, and the unit body must match the signature. If there is no U.sig, then an inferred signature file U.ui will be produced also. No evaluation takes place. During compilation, the compiled signatures of other units will be accessed if they are mentioned in U.sml.
: The declared identifiers will be reported if `verbose` is `true` (see below); otherwise compilation will be silent. In any case, compilation warnings are reported, and compilation errors abort the compilation and raise the exception `Fail` with a string argument.

`load : string -> unit`
: Evaluating `load "U"` will load and evaluate the compiled unit body from file U.uo. The resulting values are not reported, but exceptions are reported, and cause evaluation and loading to stop. If U is already loaded, then `load "U"` has no effect. If any other unit is mentioned by U but not yet loaded, then it will be loaded automatically before U.
: After loading a unit, it can be opened with `open U`. Opening it at top-level will list the identifiers declared in the unit.
: When loading U, it is checked that the signatures of units mentioned by U agree with the signatures used when compiling U, and it is checked that the signature of U has not been modified since U was compiled; these checks are necessary for type safety. The exception `Fail` is raised if these signature checks fail, or if the file containing U or a unit mentioned by U does not exist.

`loadOne : string -> unit`
: Evaluating `loadOne "U"` is similar to `load "U"`, but raises exception `Fail` if U is already loaded or if some unit mentioned by U is not yet loaded. That is, it does not automatically load any units mentioned by U. It performs the same signature checks as `load`.

`verbose : bool ref`
: Determines whether the signature inferred by a call to `compile` will be printed. The printed signature follows the syntax of Moscow ML signatures, so the output of `compile "U.sml"` can be edited to subsequently create file U.sig. The default value is `ref false`.

`quotation : bool ref`
: Determines whether quotations and antiquotations are permitted in declarations entered at top-level and in files compiled with `compile`. A quotation is a piece of text surrounded by backquote characters `` `a b c` `` and is used to embed object language phrases in ML programs; see Section 2.6

for a brief explanation of quotations. When `quotation` is `false`, the backquote character is an ordinary symbol which can be used in ML symbolic identifiers. When `quotation` is `true`, the backquote character is illegal in symbolic identifiers, and a quotation `` `a b c` `` will be recognized by the parser and evaluated to an object of type `'a General.frag list`. The default value is `ref false`.

`system : string -> int`

Evaluating `system "`*com*`"` causes the command *com* to be executed by the operating system. If a non-zero integer is returned, this must indicate that the operating system has failed to execute the command. Under MS DOS, the integer returned always equals 0.

4 Compilation units

4.1 Basic concepts

A Moscow ML program can consist of one or more *compilation units*, or *units* for short. A compilation unit consists of a *unit signature* and a *unit body*. The unit signature specifies the contents of the unit; it is an interface to the unit. The unit body declares the contents of the unit; it provides an implementation of the unit. The following analogies may be helpful:

Moscow ML	unit signature	unit body
Standard ML	signature	structure
Caml Light	module interface	module implementation
Modula-2	interface module	implementation module

The unit body is always present, whereas the signature can be omitted. When the unit signature is present, it is called the *explicit signature* to distinguish it from the signature inferred when elaborating the unit body. When present, the explicit signature must be matched by the body, and only those identifiers specified in the signature are visible outside the unit. If no signature is given, all identifiers visible at the end of the unit body are visible outside the unit.

Units are closely associated with files, as in Modula-2. The body of the unit called 'U' is defined in a file called 'U.sml', and its explicit signature (if any) in file 'U.sig'.

Files containing program text:	
U.sig	unit signature, specifications
U.sml	unit body, declarations

Files created by the compiler:	
U.ui	compiled unit signature
U.uo	compiled unit body, bytecode

4.2 Relation to Standard ML modules

A unit body U.sml without an explicit signature corresponds to a simple SML structure declaration:

```
structure U  =  struct   ... contents of file U.sml ...   end
```

An explicit unit signature U.sig corresponds to an SML signature, and the body U.sml corresponds to an SML structure with an opaque signature constraint; it is an 'abstraction' in terms of MacQueen's original proposal for SML Modules [5]:

```
signature U        =  sig      ... contents of file U.sig ...   end
abstraction U: U   =  struct   ... contents of file U.sml ...   end
```

To illustrate the difference between transparent and opaque signature constraints, consider the Standard ML declarations:

```
signature SIG = sig
    type t
    val x: t
end;
```

```
structure S: SIG = struct
    type t = int
    val x = 17
end
```

Given these declarations, the expression `S.x+33` will typecheck. Although the signature `SIG` just says that there exists a type `t`, constraining `S` with `SIG` does not hide the fact that `S.x` is actually an integer.

An opaque signature constraint, as in an `abstraction` and in Moscow ML units, does hide the true nature of `t` and `x`:

```
abstraction M: SIG = struct
    type t = int
    val x = 17
end
```

After this declaration, `M.x+33` would fail to typecheck: the type checker cannot see that `M.t` is `int` and `M.x` is an integer. Often such hiding is just what we want for software engineering purposes.

4.3 Syntax of unit signatures

Moscow ML unit signatures are very similar to Standard ML signatures as defined in [6, Figure 7]; the differences are explained below. The syntax is as follows (phrase class *uspec*):

uspec	::=	`val` *valdesc*	value specification
		`type` *typdesc*	abstract type
		`type` *typbind*	type abbreviation
		`eqtype` *typdesc*	abstract equality type
		`datatype` *datbind*	datatype
		`datatype` *datbind* `withtype` *typbind*	datatype with typbind
		`exception` *exdesc*	exception
		`local` *lspec* `in` *uspec* `end`	local specifications
			empty
		uspec $\langle ; \rangle$ *uspec*	sequential
lspec	::=	`open` $unitid_1 \cdots unitid_n$	open other units
		`type` *typbind*	type abbreviation
		`local` *lspec* `in` *lspec*	local specifications
			empty
		lspec $\langle ; \rangle$ *lspec*	sequential

Deviations from Standard ML:

1. Type abbreviations `type` *typbind* are permitted in signatures.
2. There are no structure specifications and no sharing specifications.
3. No type, value, or exception may be specified twice at top-level.
4. A `local` specification can be used only to restrict the scope of `open` specifications and type abbreviations.
5. An `open` specification can appear only inside `local`.

Restriction (2) is the most significant one. Restriction (3), and restrictions similar to (4) and (5), are imposed by the Standard ML of New Jersey implementation also.

4.4 Syntax of unit bodies

A unit body is a Core Standard ML program except that the `open` declaration is permitted, and long identifiers can be used. That is, in the definition of Core language programs [6, page 64-65] take into account only items 1 and 2.

A long identifier can refer to entities possibly declared in other units. In Moscow ML, the syntax of long identifiers is:

longid	::=	*id*	identifier
		unitid.id	qualified identifier

where *unitid* and *id* are arbitrary SML identifiers (either symbolic or alphanumeric).

A qualified identifier *unitid.id* denotes an entity *id* declared in the compilation unit *unitid*. A qualified identifier can denote either a value variable, a value constructor, an exception constructor, or a type constructor. As in Standard ML, a *longid* appearing in a defining position, such as a value variable in a pattern, cannot have a qualifier: the identifier being defined will always belong to the current unit.

An `open` declaration has the form

`open` $\mathrm{U}_1 \cdots \mathrm{U}_n$

where $\mathrm{U}_1 \cdots \mathrm{U}_n$ are names of units. The units are opened from left to right, in the order $\mathrm{U}_1 \cdots \mathrm{U}_n$. The text following an `open` U declaration can reference identifiers declared in U without explicitly specifying the name of the unit, subject to the usual scope rules of Standard ML. That is, one can use *id* instead of U.*id*.

In the interactive system, a unit must be loaded before it can be opened. In the batch compilation system, the linker links in (only) the needed declarations from opened units.

A unit body U.sml must elaborate to a structure S. If there is an explicit signature U.sig corresponding to U.sml, then the resulting structure must match the explicit signature. As in Standard ML (but in contrast to Caml Light), no reference is made to the signature while elaborating the unit body.

4.5 Matching a unit body against a signature

A unit body S matches a signature SIG under the conditions described in the Definition of Standard ML [6, Section 5.12]. Roughly, this means:

- a value specification `val` *v*:*t* must be matched by a value variable or value constructor or exception constructor *v* in S whose type generalizes *t*
- a type abbreviation `type` $t = ty$ must be matched by the same type abbreviation $t = ty$ in S
- an abstract type *t* must be matched by some type *t* in S
- an abstract equality type *t* must be matched by a type *t* in S admitting equality
- a datatype must be matched by precisely the same datatype in S
- an exception constructor *E* of type *t* must be matched by an exception constructor *E* in S whose type generalizes *t*

Moreover, to facilitate separate compilation, there are some representation constraints:

1. If the specified argument type of a value constructor (in a datatype specification) is an explicit tuple or record, then the declared argument type must be an explicit tuple or record also, and vice versa. This restriction does not apply if there is only one constructor in the datatype.
2. The order of value constructors in a datatype specification must be the same as in the matching datatype declaration.

Restriction (1) is enforced also by SML/NJ. Restriction (2) is temporary and will be lifted in future versions.

4.6 An example program consisting of three units

Here is a somewhat artificial program for working with simple arithmetic expressions. It consists of three units `Expr`, `Reduce`, and `Eval`.

This is file `Expr.sml`, the body of unit `Expr`, which defines a datatype `expr` for representing expressions, and a function `show` to display them. It has no explicit signature and therefore exports both the datatype and the function:

```
datatype expr = Cst of int | Neg of expr | Plus of expr * expr

fun show (Cst n)         = makestring n
  | show (Neg e)         = "(-" ^ show e ^ ")"
  | show (Plus (e1, e2)) = "(" ^ show e1 ^ "+" ^ show e2 ^ ")"
```

This is file `Reduce.sig`, the signature of `Reduce`, which exports a function for reduction of expressions. It mentions the type `expr` from unit `Expr`:

```
val reduce : Expr.expr -> Expr.expr
```

This is file `Reduce.sml`, the body of `Reduce`. It does not export function `negate` since it is not mentioned in the signature above:

```
local open Expr
in
    fun negate (Neg e) = e
      | negate e       = Neg e
    fun reduce (Neg (Neg e))      = e
      | reduce (Neg e)            = negate (reduce e)
      | reduce (Plus (Cst 0, e2)) = reduce e2
      | reduce (Plus (e1, Cst 0)) = reduce e1
      | reduce (Plus (e1, e2))    = Plus (reduce e1, reduce e2)
      | reduce e                  = e
end
```

This is file `Evaluate.sig`, the signature of `Evaluate`, which exports a function `eval` for evaluating expressions, and a function `test`. Note the use of '`open`' to make the type `expr` refer to `Expr.expr`:

```
local open Expr
in
    val eval : expr -> int
    val test : expr -> bool
end
```

This is file `Evaluate.sml`, the body of `Evaluate`. It mentions unit `Expr` as well as `Reduce`:

```
local open Expr
in
    fun eval (Cst n)          = n
      | eval (Neg e)          = ~ (eval e)
      | eval (Plus (e1, e2)) = eval e1 + eval e2;
    fun test e = (eval e = eval (Reduce.reduce e))
end
```

4.7 Compiling, linking, and loading units

Units can be compiled and linked using the batch compiler `mosmlc`; see Section 5. Units compiled with option `-c` can be linked together. Use `mosml -o mosmlout` to produce a linked executable bytecode file `mosmlout` which will invoke the runtime system `camlrunm`. Use `mosml -noheader -o mosmlout` to produce a linked bytecode file which can be executed by `camlrunm mosmlout`. See Section 5.2 for more options.

Units can also be compiled from and loaded into the interactive system `mosml` using the primitives `compile` and `load`; see Section 3.2 above.

4.8 Predefined and preloaded units

A number of predefined and preloaded units exist; see Section 2.5. In addition units `General` and `Misc` declare all identifiers visible at top-level in the initial environment.

5 The batch compiler

Moscow ML includes a batch compiler `mosmlc` in addition to the interactive system `mosml`. It compiles and links programs non-interactively, and can turn them into standalone executable files. The batch compiler can be invoked from a Makefile, which simplifies the (re)compilation of large programs considerably; see Section 6.

5.1 Overview

The `mosmlc` command has a command-line interface similar to that of most C compilers. It accepts several types of arguments: source files for unit bodies, source files for unit signatures, and compiled unit bodies.

- An argument ending in .sig is taken to be the name of a source file containing a unit signature. Given a file U.sig, the compiler produces a compiled signature in the file U.ui.
- An argument ending in .sml is taken to be the name of a source file containing a unit body. Given a file U.sml, the compiler produces compiled object code in the file U.uo. It also produces an inferred signature file U.ui if there is no explicit signature U.sig.
- An argument ending in .uo is taken to be the name of a compiled unit body. Such files are linked together, along with the compiled unit bodies obtained by compiling .sml arguments (if any), and the necessary Moscow ML library files, to produce a standalone executable program. The order in which .uo and .sml arguments are presented (from left to right) on the command line is significant: top-level declarations are evaluated and global identifiers are initialized in that order at run-time. In particular, file U.uo must come before any file which mentions identifiers U.*id* defined in U or which contains a declaration `open` U.

The output of the linking phase is a file containing compiled code that can be executed by the runtime system `camlrunm`. If `mosmlout` is the name of the file produced by the linking phase, the command

camlrunm mosmlout arg$_1$ arg$_2$ ... arg$_n$

executes the compiled code contained in `mosml.out`. The arguments can be accessed from the Moscow ML program via function `Mosml.argv`; see file `mosml/lib/Mosml.sig`.

MS DOS: If the output file produced by the linking phase has extension `.exe`, and option `-noheader` is not used, then the file is directly executable. Hence, an output file named `mosmlout.exe` can be executed with the command

```
mosmlout arg1 arg2 ... argn
```

The output file `mosmlout.exe` consists of a tiny executable file prepended to a linked bytecode file. The executable invokes the `camlrunm` runtime system to interpret the bytecode. As a consequence, this is not a standalone executable: it still requires `camlrunm.exe` to reside in one of the directories in the path.

Unix: The output file produced by the linking phase is directly executable (unless the `-noheader` option is used). It automatically invokes the `camlrunm` runtime system, either using a tiny executable prepended to the linked bytecode file, or using a Unix incantation such as `#!/usr/local/bin/camlrunm`. In the former case, `camlrunm` must in one of the directories in the path; in the latter case it must be in `/usr/local/bin`. To create a true stand-alone executable you may simply concatenate the runtime system with the bytecode file produced by `mosmlc -noheader`, but this adds 50–150 KB to the size of the executable, depending on your version of Unix:

```
cat /usr/local/bin/camlrunm mosmlout > mosmlbin
chmod a+x mosmlbin
```

5.2 Command-line options

The following command-line options are recognized by `mosmlc`.

`-c`
Compile only. Suppresses the linking phase of the compilation. Source code files are turned into compiled files (.ui and .uo), but no executable file is produced. This option is useful for compiling separate units.

`-files` response-file
Pass the names of files listed in file response-file to the linking phase just as if these names appeared on the command line. File names in response-file are separated by blanks (spaces, tabs, newlines) and must end either in .sml or .uo. A name U.sml appearing in the response file is equivalent to U.uo. Use this option to overcome silly limitations on the length of the command line (as in DOS).

`-g`
This option causes some information about exception names to be written at the end of the executable bytecode file. This information is required in particular by the catch-all exception handlers provided by the unit `Printexc` in the standard library.

`-i`
Causes the compiler to print the inferred signature of the unit body being compiled. A U.sig file corresponding to a U.sml file can be produced semi-automatically by redirecting the standard output of the compiler to file U.sig, and subsequently editing the file U.sig thus obtained.

`-stdlib` stdlib-directory
Specifies the standard library directory, which will be searched by the compiler and linker for the .ui and .uo files corresponding to units mentioned in the files being linked. The default standard library is set when the system is created, and is usually `${HOME}/mosml/lib` under Unix and `c:\mosml\lib` under DOS.

`-I` directory
Add the given directory to the list of directories searched for compiled signature files (.ui) and compiled object code files (.uo). By default, the current directory is searched first, then the standard library directory. Directories added with `-I` are searched after the current directory, but before the standard library directory. When several directories are added with several `-I` options on the command line, these directories are searched from left to right.

`-o` exec-file
Specify the name of the output file produced by the linker. In the absence of this option, a default name is used:

MS DOS: The default name is `mosmlout.exe`.
Unix: The default name is `a.out`.

`-P` unit-set
Determines which library units will be open at compile-time, and which library units will be available for linking at link-time.

`-P default` This provides an initial environment for the new Standard Library, compatible with future versions of SML/NJ. The units `Array`, `Char`, `List`, `String`, and `Vector` will be available for linking, and units `Char`, `List`, and `String` will be partially opened, see Section 2.1. This is the default.

`-P oldbasis` This provides an initial environment which is upwards compatible with that of the *Definition of Standard ML* and with pre-1.30 releases of Moscow ML. In particular, the functions `chr`, `explode`, `implode`, and `ord` work on strings, not characters. The new versions of these functions are still available as `Char.chr`, `Char.ord`, `String.explode`, and `String.implode`. In addition the same (new) libraries are available for linking as with `-P default`.

`-P nj93` This provides an top-level environment which is mostly compatible with that of SML/NJ 0.93. The functions `app`, `ceiling`, `chr`, `dec`, `explode`, `fold`, `hd`, `implode`, `inc`, `max`, `min`, `nth`, `nthtail`, `ord`, `ordof`, `revapp`, `revfold`, `substring`, `tl`, and `truncate` have the same type and meaning as in SML/NJ 0.93. Note that this is incompatible with future versions of SML/NJ. In addition the same (new) libraries are available for linking as with `-P default`.

`-P full` This makes all the libraries marked `F` in Section 2.5 available for linking, and partially opens the `Char`, `List`, and `String` units (see Section 2.1).

`-P none` No library units are available for linking or initially opened.

Additional library units to be linked in can be specified by listing them on the command line, or by using the `-I` directory option.

`-noheader`
: Causes the output file produced by the linker to contain only the bytecode, not preceded by the any executable code. A file `mosmlout` thus obtained can be executed only by explicitly invoking the runtime system as follows: `camlrunm mosmlout`. This option is primarily used for recompiling the system.

`-q`
: Enable the quotation/antiquotation mechanism; see Section 2.6.

`-v`
: Print the version number of the various passes of the compiler.

6 Recompilation management

Recompilation management helps the programmer recompile only what is necessary after a change to a unit signature or unit body.

Type-safe linking prevents the programmer from creating unsafe or meaningless programs. The `load` function and the batch linker ensure probabilistically type-safe linking, so it is virtually impossible to cause the system to create a type-unsafe program.

6.1 Using 'make' to manage recompilation

Consider the example program in Section 4.6 consisting of the three units Evaluate, Expr, and Reduce. Assume their source files *.sig and *.sml reside in a particular directory. Copy a Makefile stub to that directory, and change to that directory.

1. Edit the Makefile so that the names of the bytecode files Evaluate.uo, Expr.uo, and Reduce.uo appear on the line beginning with '`all:`' (see the example makefiles below).
2. Compute the dependencies among the files by executing:

   ```
   make depend
   ```

3. Recompile all those files which have not yet been compiled, or which have been modified but not yet recompiled, or which depend on modified files, by executing:

   ```
   make
   ```

Step (3) must be repeated whenever you have modified a component of the program system. Step (2) need only be repeated if the inter-dependencies of some components change, or if you add or remove an explicit signature file. Step (1) need only be repeated when you add or delete an entire unit of the program system.

Old versions of the compiled *.ui and *.uo files can be removed by executing:

```
make clean
```

The inter-dependencies are computed by a small ML program `mosmldep`, which correctly handles nested comments and strings in the source files.

6.2 An example Makefile for Unix

To use the Makefile below, first edit it so that all the required units (.uo files) appear on the line beginning with '`all:`', then proceed as explained in Section 6.1. You do not need to edit any other part of the Makefile. In particular, the dependencies following `DO NOT DELETE THIS LINE` are generated automatically when executing `make depend` (as above). A copy of the Makefile can be found in `mosml/tools/Makefile.stub`.

You will need only the Unix utility `make`.

```
MOSMLHOME=${HOME}/mosml
MOSMLTOOLS=camlrunm $(MOSMLHOME)/tools
MOSMLC=mosmlc -c

.SUFFIXES :
.SUFFIXES : .sig .sml .ui .uo

all: Evaluate.uo Expr.uo Reduce.uo

clean:
        rm -f *.ui
        rm -f *.uo
        rm -f Makefile.bak

.sig.ui:
        $(MOSMLC) $<

.sml.uo:
        $(MOSMLC) $<

depend:
        rm -f Makefile.bak
        mv Makefile Makefile.bak
$(MOSMLTOOLS)/cutdeps < Makefile.bak > Makefile
$(MOSMLTOOLS)/mosmldep >> Makefile

### DO NOT DELETE THIS LINE
Evaluate.ui: Expr.uo
Reduce.ui: Expr.uo
Evaluate.uo: Evaluate.ui Reduce.ui Expr.uo
Reduce.uo: Reduce.ui Expr.uo
```

6.3 An example Makefile for MS DOS

To use the Makefile below, first edit it so that all the required units (.uo files) appear on the line beginning with '`all:`', then proceed as explained in Section 6.1. You do not need to edit any other part of the Makefile. In particular, the dependencies following `DO NOT DELETE THIS LINE` are generated automatically when executing `make depend` (as above). A copy of this makefile can be found in `mosml\tools\makefile.stb`.

You will need a DOS version of `make`, such as that from Borland C++ version 2.0 or 3.0.

```
MOSMLHOME=c:\mosml
MOSMLTOOLS=camlrunm $(MOSMLHOME)\tools
MOSMLC=mosmlc.exe -c

all: evaluate.uo expr.uo reduce.uo

clean:
        del *.ui
        del *.uo
        del makefile.bak

.sig.ui:
        $(MOSMLC) $<

.sml.uo:
        $(MOSMLC) $<

depend:
        del makefile.bak
        ren makefile makefile.bak
        $(MOSMLTOOLS)\cutdeps < makefile.bak > makefile
        $(MOSMLTOOLS)\mosmldep >> makefile

### DO NOT DELETE THIS LINE
evaluate.ui: expr.uo
reduce.ui: expr.uo
evaluate.uo: evaluate.ui reduce.ui expr.uo
reduce.uo: reduce.ui expr.uo
```

6.4 Unit names and DOS file names

Recompilation management for DOS is essentially as for Unix, except for the usual complications that follow from the restrictions on the length of file names, and from their case-insensitivity.

Under MS DOS, filenames are all lower case and can be at most 8 characters long (plus a 3 character extension). Since file names are used as unit names, this may cause problems. We attempt to circumvent these problems as follows:

- Unit names used inside ML programs under DOS are 'normalized': the first character is made upper case (if it is a letter), all other characters are made lower case, and the unit name is truncated to eight characters. Hence a unit which resides in file `commands.sml` can be referred to as unit `Commands` inside an ML program, and can also be referred to as `CommandStructure`, etc., since normalization transforms the latter into the former.
- The following names are exceptions to this rule: `BasicIO`, `BinIO`, `CharArray`, `CharVector`, `FileSys`, `ListPair`, `OS`, `StringCvt`, `TextIO`, `Word8Array`, `Word8Vector`; they are normalized precisely as shown here. This is to accommodate the new Standard Library.
- In DOS makefiles, the file names appearing after `all:` must be all lower case and at most 8 characters long (otherwise 'make' will not work properly). For instance, the unit `CharArray` must be called `chararra` in a DOS makefile.
- A unit name given as argument to `load`, to `compile`, or to the batch compiler, is truncated and made lower case by DOS as usual, so evaluating `load "VeryLongName"` will load bytecode file `verylong.uo`.

7 A lexer generator

This section describes `mosmllex`, a lexer generator which is closely based on `camllex` from the Caml Light implementation by Xavier Leroy. This documentation is based on that of `camllex` also.

7.1 Overview

The `mosmllex` command produces a lexical analyser from a set of regular expressions with attached semantic actions, in the style of `lex`. Assume that file `lexer.lex` contains the specification of a lexical analyser. Then executing

```
mosmllex lexer.lex
```

produces a file `lexer.sml` containing Moscow ML code for the lexical analyser. This file defines one lexing function per entry point in the lexer definition. These functions have the same names as the entry points. Lexing functions take as argument a lexer buffer, and return the semantic attribute of the corresponding entry point.

Lexer buffers are an abstract data type implemented in the standard library unit `Lexing`. The functions `createLexerString` and `createLexer` from unit `Lexing` create lexer buffers that read from a character string, or any reading function, respectively.

When used in conjunction with a parser generated by `mosmlyac` (see Section 8), the semantic actions compute a value belonging to the datatype `token` defined by the generated parsing unit.

7.2 Hints on using `mosmllex`

A lexer definition must have a rule to recognize the special symbol `eof`, meaning end-of-file. In general, a lexer must be able to handle all characters that can appear in the input. This is usually achieved by having the wildcard case `_` at the very end of the lexer definition. If the lexer is to be used with e.g. DOS files, remember to provide a rule for the carriage-return symbol `\r`. Most often `\r` will be treated the same as `\n`, e.g. as whitespace.

7.3 Syntax of lexer definitions

The format of a lexer definition is as follows:

```
{ header }
rule entrypoint =
  parse regexp { action }
      | ...
      | regexp { action }
and entrypoint =
  parse ...
and ...
;
```

Comments are delimited by `(*` and `*)`, as in ML.

7.3.1 Header

The header section is arbitrary Moscow ML text enclosed in curly braces `{` and `}`. It can be omitted. If it is present, the enclosed text is copied as is at the beginning of the output file `lexer.sml`. Typically, the header section contains the `open` directives required by the actions, and possibly some auxiliary functions used in the actions.

7.3.2 Entry points

The names of the entry points must be valid ML identifiers.

7.3.3 Regular expressions

The regular expressions regexp are in the style of `lex`, but with a more ML-like syntax.

`char`
: A character constant, with a syntax similar to that of Moscow ML character constants; see Section 7.3.5. Match the denoted character.

_
: Match any character.

`eof`
: Match the end of the lexer input.

"string"
: A string constant, with a syntax similar to that of Moscow ML string constants; see Section 7.3.6. Match the denoted string.

`[` character-set `]`
: Match any single character belonging to the given character set. Valid character sets are: single character constants `` `c` ``; ranges of characters `` ` ``c_1`` ` - ` ``c_2`` ` `` (all characters between c_1 and c_2, inclusive); and the union of two or more character sets, denoted by concatenation.

`[ ^` character-set `]`
: Match any single character not belonging to the given character set.

regexp `*`
: Match the concatenation of zero or more strings that match regexp. (Repetition).

regexp `+`
: Match the concatenation of one or more strings that match regexp. (Positive repetition).

regexp `?`
: Match either the empty string, or a string matching regexp. (Option).

regexp$_1$ `|` regexp$_2$
: Match any string that matches either regexp$_1$ or regexp$_2$. (Alternative).

regexp$_1$ regexp$_2$
: Match the concatenation of two strings, the first matching regexp$_1$, the second matching regexp$_2$. (Concatenation).

`(` regexp `)`
: Match the same strings as regexp.

The operators `*` and `+` have highest precedence, followed by `?`, then concatenation, then `|` (alternative).

7.3.4 Actions

An action is an arbitrary Moscow ML expression. An action is evaluated in a context where the identifier `lexbuf` is bound to the current lexer buffer. Some typical uses of `lexbuf` in conjunction with the operations on lexer buffers (provided by the `Lexing` standard library unit) are listed below.

`Lexing.getLexeme lexbuf`
: Return the matched string.

`Lexing.getLexemeChar lexbuf n`
: Return the `n`'th character in the matched string. The first character has number 0.

`Lexing.getLexemeStart lexbuf`
: Return the absolute position in the input text of the beginning of the matched string. The first character read from the input text has position 0.

`Lexing.getLexemeEnd lexbuf`
: Return the absolute position in the input text of the end of the matched string. The first character read from the input text has position 0.

`entrypoint lexbuf`
: Here `entrypoint` is the name of another entry point in the same lexer definition. Recursively call the lexer on the given entry point. Useful for lexing nested comments, for example.

7.3.5 Character constants

A character constant in the lexer definition is delimited by ` (backquote) characters. The two backquotes enclose either a space or a printable character *c*, different from ` and \, or an escape sequence:

Sequence	Character denoted
`*c*`	the character *c*
`\\`	backslash (\)
`\``	backquote (`)
`\n`	newline (LF)
`\r`	return (CR)
`\t`	horizontal tabulation (TAB)
`\b`	backspace (BS)
`\^*c*`	the ASCII character control-*c*
`\*ddd*`	the character with ASCII code *ddd* in decimal

7.3.6 String constants

A string constant is a (possibly empty) sequence of characters delimited by " (double quote) characters.

string-literal	::=	"*strcharseq*"	non-empty string
		""	empty string
strcharseq	::=	*strchar* ⟨*strcharseq*⟩	character sequence

A string character *strchar* is either a space or a printable character *c*, different from " and \, or an escape sequence:

Sequence	Character denoted
c	the character *c*
\\	backslash (\)
\"	double quote (")
\n	newline (LF)
\r	return (CR)
\t	horizontal tabulation (TAB)
\b	backspace (BS)
\^*c*	the ASCII character control-*c*
ddd	the character with ASCII code *ddd* in decimal

8 A parser generator

This section describes `mosmlyac`, a simple parser generator which is closely based on `camlyacc` from the Caml Light implementation by Xavier Leroy; `camlyacc` in turn is closely based on Bob Corbett's public domain Berkeley `yacc`. This documentation is based on that in the Caml Light reference manual.

8.1 Overview

The `mosmlyac` command produces a parser from a context-free grammar specification with attached semantic actions, in the style of `yacc`. Assume the input file is `grammar.grm`; then executing

```
mosmlyac options grammar.grm
```

produces a file `grammar.sml` containing a Moscow ML unit with code for a parser and a file `grammar.sig` containing its interface.

The generated module defines one parsing function per entry point in the grammar. These functions have the same names as the entry points. Parsing functions take as arguments a lexical analyser (a function from lexer buffers to tokens) and a lexer buffer, and return the semantic attribute of the corresponding entry point. Lexical analyser functions are usually generated from a lexer specification by the `mosmllex` program. Lexer buffers are an abstract data type implemented in the standard library module `Lexing`. Tokens are values from the datatype `token`, defined in the interface file `grammar.sig` produced by `mosmlyac`.

8.2 Syntax of grammar definitions

A grammar definition has the following format:

```
%{
   header
%}
   declarations
%%
   rules
%%
   trailer
```

Comments in the declarations and rules sections must be enclosed in C comment delimiters `/*` and `*/`, whereas comments in the header and trailer sections must be enclosed in ML comment delimiters `(*` and `*)`.

8.2.1 Header and trailer

The header and the trailer sections consist of Moscow ML code which is copied into file `grammar.sml`. Both sections are optional. The header goes at the beginning of file `grammar.sml`, but after the `token` datatype declaration; it usually contains `open` declarations required by the semantic actions of the rules. The trailer goes at the end of file `grammar.sml`.

8.2.2 Declarations

Declarations are given one per line. They all start with a `%` sign.

`%token` symbol ... symbol
: Declare the given symbols as tokens (terminal symbols). These symbols become constructors (without arguments) in the `token` datatype.

`%token <` type `>` symbol ... symbol
: Declare the given symbols as tokens with an attached attribute of the given type. These symbols become constructors (with arguments of the given type) in the `token` datatype. The type part is an arbitrary Moscow ML type expression, but all type constructor names must be fully qualified (e.g. `Unitname.typename`) for all types except standard built-in types, even if the proper `open` declarations (e.g. `open Unitname`) were given in the header section. This is because the header is copied only to the .sml output file, not to the .sig output file, while the type part of a `%token` declaration is copied to both.

`%start` symbol ... symbol
: Declare the given symbols as entry points for the grammar. For each entry point, a parsing function with the same name is defined in the output file `grammar.sml`. Non-terminals that are not declared as entry points have no such parsing function. The start symbols must be given a type with the `%type` directive below.

`%type <` type `>` symbol ... symbol
: Specify the type of the semantic attributes for the given symbols. Every non-terminal symbol must have the type of its semantic attribute declared this way. This ensures that the generated parser is type-safe. The type part may be an arbitrary Moscow ML type expression, but all type constructor names must be fully qualified (e.g. `Unitname.typename`) for all types except standard built-in types, even if the proper `open` declaration (e.g. `open Unitname`) were given in the header section. This is

because the header is copied only to the .sml output file, not to the .sig output file, while the type part of a `%token` declaration is copied to both.

`%left` symbol ... symbol

`%right` symbol ... symbol

`%nonassoc` symbol ... symbol

Declare the precedence and associativity of the given symbols. All symbols on the same line are given the same precedence. They have higher precedence than symbols declared in previous `%left`, `%right` or `%nonassoc` lines. They have lower precedence than symbols declared in subsequent `%left`, `%right` or `%nonassoc` lines. The symbols are declared to associate to the left (`%left`), to the right (`%right`), or to be non-associative (`%nonassoc`). The symbols are usually tokens, but can also be dummy nonterminals, for use with the `%prec` directive inside the rules.

8.2.3 Rules

The format of grammar rules is as usual:

```
nonterminal :
    symbol ... symbol {  semantic-action }
  | ...
  | symbol ... symbol {  semantic-action }
;
```

Each right-hand side consists of a (possibly empty) sequence of symbols, followed by a semantic action.

The directive '`%prec` symbol' may occur among the symbols in a rule right-hand side, to specify that the rule has the same precedence and associativity as the given symbol.

Semantic actions are arbitrary Moscow ML expressions, which are evaluated to produce the semantic attribute attached to the defined nonterminal. The semantic actions can access the semantic attributes of the symbols in the right-hand side of the rule with the `$` notation: `$1` is the attribute of the first (leftmost) symbol, `$2` is the attribute of the second symbol, etc. An empty semantic action evaluates to `() : unit`.

Actions occurring in the middle of rules are not supported. Error recovery is not implemented.

8.3 Command-line options

The following command-line options are recognized by `mosmlyac`:

`-v`

Generate a description of the parsing tables and a report on conflicts resulting from ambiguities in the grammar. The description is put in file `grammar.output`.

`-b`prefix

Name the output files `prefix.sml`, `prefix.sig`, `prefix.output`, instead of using the default naming convention.

8.4 Reporting lexer and parser errors

Lexical errors (e.g. illegal symbols) and syntax errors can be reported in an intelligible way by using the `Location` module from the Moscow ML library. It provides functions to print out fragments of a source text, using location information provided by the lexer and parser. See `Location.sig` for more information.

9 Copyright and credits

Moscow ML was written by Sergei Romanenko (sergei-romanenko@refal.msk.su), Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Miusskaya Pl. 4, 125047 Moscow, Russia.

and Peter Sestoft (sestoft@dina.kvl.dk), Department of Mathematics and Physics, Royal Veterinary and Agricultural University, Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark. Much of the present work was done at the Department of Computer Science at the Technical University of Denmark, and some while visiting AT&T Bell Laboratories, Murray Hill, New Jersey, USA.

Moscow ML is based on, and would not exist without

- the CAML Light implementation by Xavier Leroy and Damien Doligez (INRIA, Rocquencourt, France);
- the ML Kit by Lars Birkedal, Nick Rothwell, Mads Tofte and David Turner (Copenhagen University, Denmark, and Edinburgh University, Scotland);
- djgpp and the go32 DOS extender by D.J. Delorie (Rochester, NH, USA);
- inspiration from the SML/NJ compiler developed at Princeton University and AT&T Bell Laboratories, New Jersey, USA.

10 How to get Moscow ML version 1.31

The Moscow ML home page is at `http://www.dina.kvl.dk/~sestoft/mosml`.

- The DOS/Windows executables are in ftp.dina.kvl.dk:pub/mosml/mos13bin.zip
- The Linux executables are in ftp://ftp.dina.kvl.dk/pub/mosml/linux-mos13bin.tar.gz
- The DOS source files are in ftp://ftp.dina.kvl.dk/pub/mosml/mos13src.zip
- The Unix source files are in ftp://ftp.dina.kvl.dk/pub/mosml/mos13src.tar.gz

You will need gcc, make, and perl to recompile Moscow ML for Unix. You will need djgpp, perl, and Borland C++ version 2.0 (or later) to recompile Moscow MLs for DOS. Djgpp and go32 can be obtained from ftp://oak.oakland.edu/SimTel/vendors/djgpp or ftp://ftp.sunet.se/pub/pc/mirror/SimTel/vendors/djgpp or any other SimTel mirror site.

11 Books and other material on Standard ML

The Definition and Commentary

- Robin Milner, Mads Tofte and Robert Harper, *The Definition of Standard ML*, MIT Press 1990, ISBN 0-262-63132-6.
- Robin Milner and Mads Tofte, *Commentary on Standard ML*, MIT Press 1991, ISBN 0-262-63137-7.

Textbooks available from publishers

- Richard Bosworth, *A Practical Course in Functional Programming Using Standard ML*, McGraw-Hill 1995, ISBN 0-07-707625-7.
- Colin Myers, Chris Clack, and Ellen Poon, *Programming with Standard ML*, Prentice Hall 1993, ISBN 0-13-722075-8.
- Lawrence C. Paulson, *ML for the Working Programmer*, Cambridge University Press 1991, ISBN 0-521-39022-2.
- Chris Reade, *Elements of Functional Programming*, Addison-Wesley 1989, ISBN 0-201-12915-9.
- Ryan Stansifer, *ML Primer*, Prentice Hall 1992, ISBN 0-13-561721-9.
- Jeffrey D. Ullman, *Elements of ML Programming*, Prentice Hall 1994, ISBN 0-13-184854-2.
- Åke Wikström, *Functional Programming Using Standard ML*, Prentice Hall 1987, ISBN 0-13-331661-0.

Texts available on the net

- Robert Harper, *Introduction to Standard ML*, LFCS Report Series ECS-LFCS-86-14, Laboratory for Foundations of Computer Science, Department of Computer Science, University of Edinburgh, November 1986 (revised 1989 by Nick Rothwell and Kevin Mitchell). Available for ftp as ftp://ftp.cs.cmu.edu/afs/cs/project/fox/mosaic/intro-notes.ps.
- Mads Tofte, *Four Lectures on Standard ML*, LFCS Report Series ECS-LFCS-89-73, Laboratory for Foundations of Computer Science, Department of Computer Science, University of Edinburgh, March 1989. Available for ftp in directory ftp://ftp.diku.dk/pub/diku/users/tofte/FourLectures/
- Mads Tofte, *Tutorial on Standard ML*, Technical Report 91/18, DIKU, Department of Computer Science, University of Copenhagen, December 1991. Available for ftp in directory ftp://ftp.diku.dk/pub/diku/users/tofte/FPCA-Tutorial/

References

[1] A. Appel et al. A new initial basis for standard ml. Technical report, AT&T Bell Laboratories, 1995. Internal draft, February 6.

[2] L. Birkedal, N. Rothwell, M. Tofte, and D.N. Turner. The ML Kit. Technical Report 93/14, DIKU, University of Copenhagen, Denmark, 1993.

[3] X. Leroy. The Zinc experiment: An economical implementation of the ML language. Rapport Technique 117, INRIA Rocquencourt, France, 1990. Available as ftp.inria.fr:/INRIA/Projects/cristal/Xavier.Leroy/economical-ML-implementation.ps.gz.

[4] X. Leroy. *The Caml Light system, release 0.6. Documentation and user's manual.* INRIA, France, September 1993. Available at ftp.inria.fr:lang/caml-light.

[5] D.B. MacQueen. Modules for Standard ML. In *Proc. 1984 ACM Conf. on LISP and Functional Programming*, pages 198–207. ACM Press, 1984.

[6] R. Milner, M. Tofte, and R. Harper. *The Definition of Standard ML*. MIT Press, 1990.

[7] S. Romanenko and P. Sestoft. *Moscow ML Language Overview*, version 1.30 edition, March 1995.