INF564 – Compilation

Jean-Christophe Filliâtre

abstract syntax, semantics
how to define the meaning of programs?

most of the time, we are satisfied with an informal description, in natural language (ISO norm, standard, reference book, etc.)

yet it is imprecise, sometimes even ambiguous
The Java programming language guarantees that the operands of operators appear to be evaluated in a specific evaluation order, namely, from left to right.

It is recommended that code not rely crucially on this specification.
**formal semantics** gives a mathematical characterization of the computations defined by a program

useful to make tools (interpreters, compilers, etc.)

necessary to reason about programs
what is a program?

as a syntactic object (sequence of characters),
it is too complex to apprehend

that’s why we switch to abstract syntax
source
↓
lexical analysis
↓
stream of tokens
↓
parsing
↓
abstract syntax tree
↓
semantic analysis
↓
abstract syntax + symbol table
↓

code production
↓
assembly code
↓
assembler (as)
↓
machine language
↓
linking (ld)
↓
executable
the texts

\[ 2 \times (x + 1) \]

and

\[ (2 \times ((x) + 1)) \]

and

\[ 2 \times /* I \ double */ (x + 1) \]

all map to the same \textit{abstract syntax tree}

\[
\begin{array}{c}
  \times \\
  \bigg/ \bigg/ \\
  2 \bigg/ \bigg/ + \\
  \bigg/ \bigg/ \\
  x \bigg/ \bigg/ 1 \\
\end{array}
\]
we define an abstract syntax using a **grammar**

$$e ::= c \quad \text{constant}$$

$$\mid x \quad \text{variable}$$

$$\mid e + e \quad \text{addition}$$

$$\mid e \times e \quad \text{multiplication}$$

$$\mid \ldots$$

reads “an expression, noted $e$, is

- either a constant $c$,
- either a variable $x$,
- either the addition of two expressions,
- etc.”
notation $e_1 + e_2$ of the abstract syntax borrows the symbol of the concrete syntax

but we could have picked something else, e.g. $Add(e_1, e_2)$, $+(e_1, e_2)$, etc.
we use classes to build abstract syntax trees, as follows:

```java
enum Binop { Add, Mul, ... }

abstract class Expr {}

class Cte extends Expr { int n; }
class Var extends Expr { String x; }
class Bin extends Expr { Binop op; Expr e1, e2; }
...
```

(expression constructors are omitted)

expression 2 * (x + 1) is then represented as

```java
new Bin(Mul, new Cte(2), new Bin(Add, new Var("x"), new Cte(1)))
```
we use algebraic data types to abstract syntax trees, as follows:

```ocaml
type binop = Add | Mul | ...

type expression =
    | Cte of int
    | Var of string
    | Bin of binop * expression * expression
    | ...
```

expression $2 \times (x + 1)$ is then represented as

```
Bin (Mul, Cte 2, Bin (Add, Var "x", Cte 1))
```
there is no constructor for parentheses in abstract syntax

in \textbf{concrete} syntax $2 \times (x + 1)$, parentheses are used to build this tree

\[
\begin{array}{c}
\times \\
\bigg/ \bigg/ \\
2 \bigg/ + \\
\bigg/ \bigg/ \\
x \bigg/ 1
\end{array}
\]

rather than this one

\[
\begin{array}{c}
+ \\
\bigg/ \bigg/ \\
\times 1 \\
\bigg/ \bigg/ \\
2 \bigg/ x
\end{array}
\]

(the lecture on parsing will explain how)
we call **syntactic sugar** a construct of concrete syntax that does not exist in abstract syntax

it is thus translated in terms of other constructs of abstract syntax (typically during parsing)

effects:
- in C, expression `a[i]` is syntactic sugar for `*(a+i)`
- in Java, expression `x -> {...}` is sugar for the construction of an object in some anonymous class that implements `Function`
- in OCaml, expression `[e_1; e_2; ...; e_n]` is sugar for `e_1 :: e_2 :: ... :: e_n :: []`
formal semantics is defined over abstract syntax

there are many approaches

• axiomatic semantics
• denotational semantics
• semantics by translation
• operational semantics
also called **Floyd-Hoare logic**
(Robert Floyd, *Assigning meanings to programs*, 1967
Tony Hoare, *An axiomatic basis for computer programming*, 1969)
defines programs by means of their properties; we introduce a triple

\[ \{P\} \ i \ \{Q\} \]

meaning “if formula $P$ holds before the execution of statement $i$, then formula $Q$ holds after the execution”

dexample:

\[ \{x \geq 0\} \ x := x + 1 \ \{x > 0\} \]

dexample of rule:

\[ \{P[x \leftarrow E]\} \ x := E \ \{P(x)\} \]
denotational semantics maps each program expression $e$ to its denotation $\llbracket e \rrbracket$, a mathematical object that represents the computation denoted by $e$.

Example: arithmetic expressions with a single variable $x$

$$e ::= x \mid n \mid e + e \mid e \ast e \mid \ldots$$

The denotation is a function that maps the value of $x$ to the value of the expression:

$$\llbracket x \rrbracket = x \mapsto x$$
$$\llbracket n \rrbracket = x \mapsto n$$
$$\llbracket e_1 + e_2 \rrbracket = x \mapsto \llbracket e_1 \rrbracket(x) + \llbracket e_2 \rrbracket(x)$$
$$\llbracket e_1 \ast e_2 \rrbracket = x \mapsto \llbracket e_1 \rrbracket(x) \times \llbracket e_2 \rrbracket(x)$$
(also called Strachey semantics)

we can define the semantics of a language by means of its translation to another language for which the semantics is already defined
an esoteric language whose syntax consists of 8 characters and whose semantics is defined by translation to the C language

<table>
<thead>
<tr>
<th>command</th>
<th>translation to C</th>
</tr>
</thead>
</table>
| (prelude) | char array[30000] = {0};
|          | char *ptr = array; |
| >       | ++ptr;           |
| <       | --ptr;           |
| +       | +++*ptr;         |
| -       | --*ptr;          |
| .       | putchar(*ptr);   |
| ,       | *ptr = getchar(); |
| [       | while (*ptr) {   |
| ]       | }                |
**operational semantics** describes the sequence of elementary computations from the expression to its outcome (its value)

it operates directly over abstract syntax

two kinds of operational semantics

- “natural semantics” or “big steps”

\[
e \rightarrow v
\]

- “reduction semantics” or “small steps”

\[
e \rightarrow e_1 \rightarrow e_2 \rightarrow \cdots \rightarrow v
\]
let us illustrate big-step operational semantics on a minimal language

\[
e ::= \text{expression} \\
| \quad c \quad \text{constant} \\
| \quad x \quad \text{variable} \\
| \quad e \ op \ e \quad \text{binary operator (+, <...)}
\]

\[
c ::= \text{constant} \\
| \quad n \quad \text{integer constant (-17, 42...)} \\
| \quad b \quad \text{Boolean constant (true, false)}
\]
\[ s ::= \text{statement} \]
\[ \quad \mid x \leftarrow e \quad \text{assignment} \]
\[ \quad \mid \text{if } e \text{ then } s \text{ else } s \quad \text{conditional} \]
\[ \quad \mid \text{while } e \text{ do } s \quad \text{loop} \]
\[ \quad \mid s; s \quad \text{sequence} \]
\[ \quad \mid \text{skip} \quad \text{do nothing} \]
\begin{verbatim}
a ← 0;
b ← 1;
while b < 1000 do
    b ← a + b;
    a ← b − a
\end{verbatim}
big steps operational semantics of \textbf{WHILE}

we seek to define a relation between some expression $e$ and a \textbf{value} \(v\)

\[
e \rightarrow v
\]

here, values are limited to integer and Boolean constants

\[
v ::= \text{value} \\
| n \quad \text{integer value} \\
| b \quad \text{Boolean value}
\]

caveat: with most languages, values do not coincide with constants
in Java (or Python, OCaml, etc.), a value may be an address, even if we do not have addresses among the literal constants of the language.

```java
int[] a = new int[4];
...
int[] b = a;
b[2] = 42;
...
```

(more about this in lecture 5)
the value of a variable is given by an environment $E$ (a function from variables to values)

we are going to define a relation

$$E, e \rightarrow v$$

that reads “in environment $E$, expression $e$ has value $v$”
in environment

\[ E = \{ a \mapsto 34, \ b \mapsto 55 \} \]

the expression

\[ a + b \]

has value

\[ 89 \]

which we write

\[ E, a + b \rightarrow 89 \]
a relation may be defined as the **smallest relation** satisfying a set of rules with no premises (axioms) written

\[
\begin{array}{c}
P \\
\end{array}
\]

and a set of rules with premises written

\[
\begin{array}{c}
P_1 & P_2 & \ldots & P_n \\
\end{array}
\]

\[
\begin{array}{c}
P \\
\end{array}
\]

this is called **inference rules**
we can define the relation \( \text{Even}(n) \) with two rules

\[
\frac{\text{Even}(0)}{\text{Even}(n)} \quad \text{et} \quad \frac{\text{Even}(n)}{\text{Even}(n+2)}
\]

that reads as follows

on the one hand \( \text{Even}(0) \)
on the other hand \( \forall n. \text{Even}(n) \Rightarrow \text{Even}(n+2) \)

the smallest relation satisfying these two properties coincide with the property “\( n \) is an even natural number”:

• even natural numbers are included, by induction
• if odd numbers were included, we could remove the smallest
A derivation is a tree whose internal nodes are rules with premises and whose leaves are axioms.

Example:

```
  Even(0)
  Even(2)
  Even(4)
```

The set of derivations characterizes the smallest relation satisfying the inference rules.
the relation $E, e \rightarrow v$ is defined by the following inference rules:

$$E, n \rightarrow n \quad E, b \rightarrow b$$

$$x \in \text{dom}(E) \quad \frac{}{E, x \rightarrow E(x)}$$

$$E, e_1 \rightarrow n_1 \quad E, e_2 \rightarrow n_2 \quad n = n_1 + n_2 \quad \frac{}{E, e_1 + e_2 \rightarrow n}$$

etc.
with \( E = \{ a \mapsto 34, \ b \mapsto 55 \} \), we have

\[
\frac{a \in \text{dom}(E)}{E, a \rightarrow 34} \quad \frac{b \in \text{dom}(E)}{E, b \rightarrow 55} \quad 89 = 34 + 55
\]

\[
E, a + b \rightarrow 89
\]

note: one can see such a tree as a proof
A statement may modify the value of some variables (through assignments). To define the semantics of a statement $s$, we thus introduce the relation

$$ E, s \rightarrow E' $$

that reads “in environment $E$, the evaluation of statement $s$ terminates and leads to environment $E'$”
semantics of statements

\[
\begin{align*}
E, \text{skip} & \rightarrow E \\
E, s_1 \rightarrow E_1 & \quad E_1, s_2 \rightarrow E_2 \\
E, s_1 ; s_2 & \rightarrow E_2 \\
E, e \rightarrow v & \\
E, x \leftarrow e & \rightarrow E\{x \leftarrow v\} \\
E, e \rightarrow \text{true} & \quad E, s_1 \rightarrow E_1 \\
E, \text{if } e \text{ then } s_1 \text{ else } s_2 & \rightarrow E_1 \\
E, e \rightarrow \text{false} & \quad E, s_2 \rightarrow E_2 \\
E, \text{if } e \text{ then } s_1 \text{ else } s_2 & \rightarrow E_2 \\
E, e \rightarrow \text{true} & \quad E, s \rightarrow E_1 \\
E_1, \text{while } e \text{ do } s & \rightarrow E_2 \\
E, \text{while } e \text{ do } s & \rightarrow E_2 \\
E, e \rightarrow \text{false} & \\
E, \text{while } e \text{ do } s & \rightarrow E
\end{align*}
\]
with $E = \{ a \mapsto 21 \}$, we have

$$
\frac{E, a \mapsto 21 \quad E, 0 \mapsto 0}{E, a > 0 \mapsto \text{true}} \quad \frac{E, 2 \mapsto 2 \quad E, a \mapsto 21}{E, 2 \times a \mapsto 42} \quad \frac{E, a \leftarrow 2 \times a \mapsto \{ a \mapsto 42 \}}{E, \text{if } a > 0 \text{ then } a \leftarrow 2 \times a \text{ else skip} \mapsto \{ a \mapsto 42 \}}
$$
there are expressions $e$ for which there is no value $v$ such that $E, e \leadsto v$

example: $1 + \text{true}$

similarly, there are statements $s$ for which there is no evaluation $E, s \leadsto E'$

example: while true do skip
to establish a property of a relation defined by a set of inference rules, one can reason by **structural induction** on the derivation, i.e. one can use the induction hypothesis on any sub-derivation.

equivalently, one can say that we perform an induction over the height of the derivation.

in practice, we proceed by induction on the derivation and **by case** on the last rule of the derivation.
Proposition (evaluation is deterministic)

If $E, e \rightarrow v$ and $E, e \rightarrow v'$ then $v = v'$.

by induction over the derivations of $E, e \rightarrow v$ and $E, e \rightarrow v'$
case of an addition $e = e_1 + e_2$

\[ (D_1) \quad (D_2) \]
\[
\begin{array}{c}
E, e_1 \rightarrow n_1 \\
E, e_2 \rightarrow n_2 \\
\hline
E, e_1 + e_2 \rightarrow v
\end{array}
\]

\[ (D'_1) \quad (D'_2) \]
\[
\begin{array}{c}
E, e_1 \rightarrow n'_1 \\
E, e_2 \rightarrow n'_2 \\
\hline
E, e_1 + e_2 \rightarrow v'
\end{array}
\]

with $v = v_1 + n_2$ et $v' = n_1 + n_2$
by IH we have $n_1 = n'_1$ and $n_2 = n'_2$ and thus $v = v'$

(other cases are similar or simpler)
Proposition (evaluation is deterministic)

If $E,s \rightarrow E'$ and $E,s \rightarrow E''$ then $E' = E''$.

exercise: do this proof

remark: in the case of rule

$$
\frac{E,e \rightarrow \text{true} \quad E,s \rightarrow E_1 \quad E_1,\text{while } e \text{ do } s \rightarrow E_2}{E,\text{while } e \text{ do } s \rightarrow E_2}
$$

it is clear that induction is performed on the size of the derivation and not on the size of the statement (which does not decrease)
an evaluation relation is not necessarily deterministic

example: we add a primitive \textit{random} to draw an integer 0 or 1 at random, with the rule

\[
\frac{0 \leq n < 2}{E, \text{random} \rightarrow n}
\]

then we have \( E, \text{random} \rightarrow 0 \) as well as \( E, \text{random} \rightarrow 1 \)
we can code an interpreter following the rules of the natural semantics

let’s do it in Java
as explained earlier

```java
enum Binop { Add, ... }

abstract class Expr {}
class Ecte extends Expr { Value v; }
class Evar extends Expr { String x; }
class Ebin extends Expr { Binop op; Expr e1, e2; }

abstract class Value {}
class Vint extends Value { int n; }
class Vbool extends Value { boolean b; }
```

(constructors are omitted)
abstract syntax

similarly for statements

```
abstract class Stmt {}
class Sskip extends Stmt {}
class Sassign extends Stmt { String x; Expr e; }
class Sif extends Stmt { Expr e; Stmt s1, s2; }
class Swhile extends Stmt { Expr e; Stmt s; }
class Sseq extends Stmt { Stmt s1, s2; }
```
let's start with relation

\[ E, e \rightarrow v \]

the environment \( E \) is represented by a class

```java
class Environment {
    HashMap<String, Value> vars = new HashMap<>();
}
```
one solution is to declare a method

```java
abstract class Expr {
    abstract Value eval(Environment env);
}
```

and then to define it within any sub-class
evaluation of an expression

| $E, n \rightarrow n$ | $E, b \rightarrow b$ |

```java
class Ecte extends Expr {
    Value eval(Environment env) {
        return v;
    }
}
class Evar extends Expr {
    Value eval(Environment env) {
        Value v = env.vars.get(x);
        if (v == null)
            throw new Error("unbound variable " + x);
        return v;
    }
}
```
Evaluation of an expression:

\[
\frac{E, e_1 \rightarrow n_1 \quad E, e_2 \rightarrow n_2 \quad n = n_1 + n_2}{E, e_1 + e_2 \rightarrow n}
\]

etc.

class Ebin extends Expr {
    Value eval(Environment env) {
        Value v1 = e1.eval(env), v2 = e2.eval(env);
        switch (op) {
            case Add:
                return new Vint(v1.asInt() + v2.asInt());
            ...
        }
    }
}
method `asInt` (in class `Value`) checks that the value is an integer and returns it otherwise, `asInt` raises an exception

consequently, our interpreter fails **dynamically** on an expression such as `1+true`

we could have detected this error **statically** with typing (see lecture 4)

\[
\text{statically} \quad = \quad \text{at compile time}
\]
\[
\text{dynamically} \quad = \quad \text{during execution}
\]
we proceed similarly for statements by adding a method in class Stmt

```java
abstract class Stmt {
    abstract void eval(Environment env);
}
```

that we define within any sub-class

eval returns nothing, but it mutates the environment
evaluation of a statement

\[ E, \text{skip} \rightarrow E \]

class Sskip extends Stmt {
    void eval(Environment env) {}
}

\[
\frac{E, s_1 \rightarrow E_1 \quad E_1, s_2 \rightarrow E_2}{E, s_1; s_2 \rightarrow E_2}
\]

class Sseq extends Stmt {
    void eval(Environment env) {
        s1.eval(env);
        s2.eval(env);
    }
}
evaluation of a statement

\[
E, e \rightarrow v
\]

\[
E, x \leftarrow e \rightarrow E\{x \mapsto v\}
\]

class Sassign extends Stmt {
    void eval(Environment env) {
        env.vars.put(x, e.eval(env));
    }
}

(the environment is a mutable data structure)
evaluation of a statement

\[ \begin{align*}
E, e \to true & \quad E, s_1 \to E_1 \\
E, if \; e \; then \; s_1 \; else \; s_2 \to E_1 \\
E, e \to false & \quad E, s_2 \to E_2 \\
E, if \; e \; then \; s_1 \; else \; s_2 \to E_2
\end{align*} \]

class Sif extends Stmt {
    void eval(Environment env) {
        if (e.eval(env).asBoolean())
            s1.eval(env);
        else
            s2.eval(env);
    }
}
evaluation of a statement

\[
\begin{align*}
E, e \rightarrow \text{true} & \quad E, s \rightarrow E_1 & \quad E_1, \text{while } e \text{ do } s \rightarrow E_2 \\
\hline
E, \text{while } e \text{ do } s \rightarrow E_2
\end{align*}
\]

\[
\begin{align*}
E, e \rightarrow \text{false} & \\
\hline
E, \text{while } e \text{ do } s \rightarrow E
\end{align*}
\]

class Swhile extends Stmt {
    void eval(Environment env) {
        while (e.eval(env).asBoolean())
            s.eval(env);
    }
}
we can do the same in OCaml

pattern matching plays the role of dynamic methods

```
let rec eval env = function
  | Ecte v ->
    v
  | Evar x ->
    (try Hashtbl.find env x
     with Not_found -> failwith ("unbound variable" ^ x))
  | Ebin (op, e1, e2) ->
    (match op, eval env e1, eval env e2 with
     | Add, Vint n1, Vint n2 -> Vint (n1 + n2)
     | ...)
  | _ -> failwith "illegal operands")
```
brief comparison functional/object programming

what distinguishes

type expr = Cte of value | Evar of string | ...

abstract class Expr {...} class Ecte extends Expr {...}

in OCaml, the code of eval is a single function and it covers all cases

in Java, it is scattered in all classes
A brief comparison of functional/object programming:

<table>
<thead>
<tr>
<th></th>
<th>horizontal extension</th>
<th>vertical extension</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>= adding a case</td>
<td>= adding a function</td>
</tr>
<tr>
<td>Java</td>
<td>easy</td>
<td>painful</td>
</tr>
<tr>
<td></td>
<td>(one file)</td>
<td>(several files)</td>
</tr>
<tr>
<td>OCaml</td>
<td>painful</td>
<td>easy</td>
</tr>
<tr>
<td></td>
<td>(several files)</td>
<td>(one file)</td>
</tr>
</tbody>
</table>
another way of writing the Java code

the Java code may be organized differently, with

• classes for the abstract syntax on one side,
• a class for the interpreter on the other side

to do that, we can use the **visitor pattern**
we start by introducing an interface for the interpreter

```java
interface Interpreter {
    Value interp(Ecte e);
    Value interp(Evar e);
    Value interp(Ebin e);
}
```

note: we use Java’s **overloading** to give all these methods the same name
in class Expr, we provide a method accept to apply the interpreter

```java
abstract class Expr {
    abstract Value accept(Interpreter i);
}
class Ecte extends Expr {
    Value accept(Interpreter i) { return i.interp(this); }
}
class Evar extends Expr {
    Value accept(Interpreter i) { return i.interp(this); }
}
class Ebin extends Expr {
    Value accept(Interpreter i) { return i.interp(this); }
}
```

this is the only intrusion in the classes of abstract syntax
finally, we can code the interpreter in a separate class, that implements interface Interpreter

```java
class Interp implements Interpreter {
    Environment env = new Environment();
    Value interp(Ecte e) {
        return e.v;
    }
    Value interp(Ebin e) {
        Value v1 = e.e1.accept(this), v2 = e.e2.accept(this);
        switch (e.op) {
            case Add:
                return new Vint(v1.asInt() + v2.asInt());
            ...
        }
    }
    ...
}
```
the interface Interpreter is specific to our needs

we could instead provide a general-purpose visitor

```
interface Visitor {
    void visit(Ecte e);
    void visit(Evar e);
    void visit(Ebin e);
}
```

so that we can use it for other purposes, e.g. printing

methods visit return nothing, but this is not an issue (see this week’s lab)
weaknesses of natural semantics

natural semantics makes no distinction between programs that crash, such as

\[ 1 + \text{true} \]

and programs whose evaluation does not terminate, such as

\[ \text{while true do skip} \]
small-step operational semantics remedies this by introducing a notion of elementary computation $E_1, s_1 \rightarrow E_2, s_2$, which we iterate

then we can distinguish

1. successful termination

$$E, s \rightarrow E_1, s_1 \rightarrow E_2, s_2 \rightarrow \cdots \rightarrow E', \text{skip}$$

2. evaluation stuck on $E_n, s_n$ with $s_n \neq \text{skip}$

$$E, s \rightarrow E_1, s_1 \rightarrow E_2, s_2 \rightarrow \cdots \rightarrow E_n, s_n$$

3. non-terminating evaluation

$$E, s \rightarrow E_1, s_1 \rightarrow E_2, s_2 \rightarrow \cdots$$
we can keep our big-step semantics for expressions, since expressions always terminate

but for a more complex language (say, with function calls in expressions), we would use small-step semantics for expressions as well
small-step operational semantics for \texttt{WHILE}

\[
\begin{align*}
E, e \rightarrow v \\
E, x \leftarrow e \rightarrow E\{x \leftarrow v\}, \text{skip}
\end{align*}
\]

\[
\begin{align*}
E, \text{skip}; s \rightarrow E, s \\
E, s_1 \rightarrow E_1, s'_1 \\
E, s_1; s_2 \rightarrow E_1, s'_1; s_2
\end{align*}
\]

\[
\begin{align*}
E, e \rightarrow \text{true} \\
E, \text{if } e \text{ then } s_1 \text{ else } s_2 \rightarrow E, s_1
\end{align*}
\]

\[
\begin{align*}
E, e \rightarrow \text{false} \\
E, \text{if } e \text{ then } s_1 \text{ else } s_2 \rightarrow E, s_2
\end{align*}
\]

\[
\begin{align*}
E, e \rightarrow \text{true} \\
E, \text{while } e \text{ do } s \rightarrow E, s; \text{while } e \text{ do } s
\end{align*}
\]

\[
\begin{align*}
E, e \rightarrow \text{false} \\
E, \text{while } e \text{ do } s \rightarrow E, \text{skip}
\end{align*}
\]
\{\}, x \leftarrow 40; \text{while } x < 42 \text{ do } x \leftarrow x + 1
\rightarrow \{x \mapsto 40\}, \text{skip; while } x < 42 \text{ do } x \leftarrow x + 1
\rightarrow \{x \mapsto 40\}, \text{while } x < 42 \text{ do } x \leftarrow x + 1
\rightarrow \{x \mapsto 40\}, x \leftarrow x + 1; \text{while } x < 42 \text{ do } x \leftarrow x + 1
\rightarrow \{x \mapsto 41\}, \text{skip; while } x < 42 \text{ do } x \leftarrow x + 1
\rightarrow \{x \mapsto 41\}, \text{while } x < 42 \text{ do } x \leftarrow x + 1
\rightarrow \{x \mapsto 41\}, x \leftarrow x + 1; \text{while } x < 42 \text{ do } x \leftarrow x + 1
\rightarrow \{x \mapsto 42\}, \text{skip; while } x < 42 \text{ do } x \leftarrow x + 1
\rightarrow \{x \mapsto 42\}, \text{while } x < 42 \text{ do } x \leftarrow x + 1
\rightarrow \{x \mapsto 42\}, \text{skip}
we could replace the two rules for \texttt{while} with the following rule

$$E, \texttt{while} \ e \ \texttt{do} \ s \rightarrow E, \texttt{if} \ e \ \texttt{then} \ s; \ \texttt{while} \ e \ \texttt{do} \ s \ \texttt{else} \ \texttt{skip}$$
**Proposition (equivalence of the two semantics)**

*The two operational semantics are equivalent on programs whose evaluation terminate, i.e.*

\[ E, s \rightarrow E' \quad \text{if and only if} \quad E, s \rightarrow^* E', \text{skip} \]

*(where \( \rightarrow^* \) is the reflexive transitive closure of \( \rightarrow \)).*
Proposition (big steps imply small steps)

If $E, s \rightarrow E'$, then $E, s \rightarrow^{*} E', \text{skip}$.

by induction on the derivation $E, s \rightarrow E'$ and by case on the last rule

• case of $s_1; s_2$

\[
\begin{align*}
E, s_1 &\rightarrow E_1 \\
E_1, s_2 &\rightarrow E_2 \\
\hline
E, s_1; s_2 &\rightarrow E_2
\end{align*}
\]

then $E, s_1 \rightarrow^{*} E_1, \text{skip}$ by IH

consequently,

\[
\begin{align*}
E, s_1; s_2 &\rightarrow^{*} E_1, \text{skip}; s_2 \quad \text{(small steps for ;)} \\
&\rightarrow E_1, s_2 \\
&\rightarrow^{*} E_2, \text{skip} \quad \text{(IH)}
\end{align*}
\]
• case of \texttt{while e do s}

if

\[
\begin{align*}
E, e & \rightarrow \text{true} & E, s & \rightarrow E_1 & E_1, \text{while e do s} & \rightarrow E_2 \\
\end{align*}
\]

then

\[
\begin{align*}
E, \text{while e do s} & \rightarrow E, s; \text{while e do s} \\
& \rightarrow^* E_1, \text{skip}; \text{while e do s} & (\text{IH + rule ;}) \\
& \rightarrow E_1, \text{while e do s} \\
& \rightarrow^* E_2, \text{skip} & (\text{IH})
\end{align*}
\]

exercise: do the other cases
Lemma

If $E_1, s_1 \rightarrow E_2, s_2 \rightarrow E'$, then $E_1, s_1 \rightarrow E'$.

by induction over $\rightarrow$

- case $s_1 = u_1; v_1$

  - case $u_1 = \text{skip}$
    
    we have $E_1, \text{skip}; v_1 \rightarrow E_1, v_1 \rightarrow E'$ and thus

    \[
    \frac{E_1, \text{skip} \rightarrow E_1 \quad E_1, v_1 \rightarrow E'}{E_1, \text{skip}; v_1 \rightarrow E'}
    \]

  - case $u_1 \neq \text{skip}$
    
    we have $E_1, u_1; v_1 \rightarrow E_2, u_2; v_1 \rightarrow E'$ that is $E_1, u_1 \rightarrow E_2, u_2$ and

    \[
    \frac{E_2, u_2 \rightarrow E_2' \quad E_2', v_1 \rightarrow E'}{E_2, u_2; v_1 \rightarrow E'}
    \]

    by IH we deduce

    \[
    \frac{E_1, u_1 \rightarrow E_2' \quad E_2', v_1 \rightarrow E'}{E_1, u_1; v_1 \rightarrow E'}
    \]

(dо the other cases)
we deduce

Proposition (small steps imply big steps)

Si $E, s \rightarrow^* E', \text{skip}$, alors $E, s \rightarrow E'$.

proof: we have

$$E, s \rightarrow E_1, s_1 \rightarrow E_2, s_2 \rightarrow \cdots \rightarrow E_n, s_n \rightarrow E', \text{skip}$$

but $E', \text{skip} \rightarrow E'$ so $E_n, s_n \rightarrow E'$ by the lemma above, then $E_{n-1}, s_{n-1} \rightarrow E'$ by the same lemma, etc., until we get $E, s \rightarrow E'$

(induction on $n$, the number of steps)
the lecture notes also contain the operational semantics for Mini-ML

\[
e ::= x \quad \text{identifier} \\
| c \quad \text{constant (1, 2, \ldots, true, \ldots)} \\
| op \quad \text{primitive (+, \times, fst, \ldots)} \\
| \text{fun } x \to e \quad \text{anonymous function} \\
| e \ e \quad \text{application} \\
| (e, e) \quad \text{pair} \\
| \text{let } x = e \ \text{in } e \quad \text{local binding}
\]

(section 2.2, page 20)
application

correctness of a compiler
a compiler must respect the semantics

if the input language is equipped with a semantics $\rightarrow_s$ and the target language with a semantics $\rightarrow_m$, and if some expression $e$ is compiled to $C(e)$ the we must have “a commutative diagram”:

$$
\begin{array}{ccc}
  e & \xrightarrow{\ast} & v \\
  \downarrow & & \approx \\
  C(e) & \xrightarrow{\ast} & v'
\end{array}
$$

where $v \approx v'$ states that values $v$ and $v'$ coincide
let us consider arithmetic expressions with no variables

\[ e ::= n \mid e + e \]

and let us show the correctness of a very simple compiler to x86-64 that uses the stack to store intermediate computations
we set a small-step semantics for the input language

\[
\begin{align*}
n &= n_1 + n_2 \\
n_1 + n_2 &\rightarrow n \\
e_1 &\rightarrow e'_1 \\
e_1 + e_2 &\rightarrow e'_1 + e_2 \\
e_2 &\rightarrow e'_2 \\
n_1 + e_2 &\rightarrow n_1 + e'_2
\end{align*}
\]
simply, we set a small-step semantics for the target language

\[ m ::= \text{movq } n, r \]
\[ \quad | \quad \text{addq } n, r \mid \text{addq } r, r \]
\[ \quad | \quad \text{movq } (r), r \mid \text{movq } r, (r) | \]
\[ r ::= \%rdi \mid \%rsi \mid \%rsp \]

a state gathers the values of registers, \( R \), and the contents of the memory, \( M \)

\[ R ::= \{ \%rdi \mapsto n; \%rsi \mapsto n; \%rsp \mapsto n \} \]
\[ M ::= \mathbb{N} \rightarrow \mathbb{Z} \]

we then define the semantics of an instruction \( m \) using a relation

\[ R, M, m \xrightarrow{m} R', M' \]
the relation $R, M, m \xrightarrow{m} R', M'$ is defined as follows:

\[
\begin{align*}
    R, M, \text{movq } n, r & \xrightarrow{m} R\{r \mapsto n\}, M \\
    R, M, \text{addq } n, r & \xrightarrow{m} R\{r \mapsto R(r) + n\}, M \\
    R, M, \text{addq } r_1, r_2 & \xrightarrow{m} R\{r_2 \mapsto R(r_1) + R(r_2)\}, M \\
    R, M, \text{movq } (r_1), r_2 & \xrightarrow{m} R\{r_2 \mapsto M(R(r_1))\}, M \\
    R, M, \text{movq } r_1, (r_2) & \xrightarrow{m} R, M\{R(r_2) \mapsto R(r_1)\}
\end{align*}
\]
the final value of an expression is stored in %rdi

\[
\text{code}(n) = \text{movq} \; n, %rdi
\]

\[
\text{code}(e_1 + e_2) = \text{code}(e_1) \\
\quad \text{addq} \; -8, %rsp \\
\quad \text{movq} \; %rdi, (%rsp) \\
\quad \text{code}(e_2) \\
\quad \text{movq} \; (%rsp), %rsi \\
\quad \text{addq} \; 8, %rsp \\
\quad \text{addq} \; %rsi, %rdi
\]
we seek to prove that if
\[ e \xrightarrow{\star} n \]
and if
\[ R, M, \text{code}(e) \xrightarrow{m}^* R', M' \]
then \( R'(\%rdi) = n \)

we proceed by structural induction on \( e \)
we show a stronger property (an **invariant**), namely

if $e \xrightarrow{*} n$ and $R, M, code(e) \xrightarrow{m} R', M'$ then

$$\begin{align*}
R'(\%rdi) &= n \\
R'(\%rsp) &= R(\%rsp) \\
\forall a \geq R(\%rsp), \; M'(a) &= M(a)
\end{align*}$$
• case $e = n$

we have $e \xrightarrow{*} n$ and $\text{code}(e) = \text{movq} \ $n, %rdi$ and the result is immediate

• case $e = e_1 + e_2$

we have $e \xrightarrow{*} n_1 + e_2 \xrightarrow{*} n_1 + n_2$ with $e_1 \xrightarrow{*} n_1$ and $e_2 \xrightarrow{*} n_2$

thus we can invoke the induction hypothesis on $e_1$ and $e_2$
<table>
<thead>
<tr>
<th></th>
<th>$R, M$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>code($e_1$)</strong></td>
<td>$R_1, M_1$</td>
</tr>
<tr>
<td></td>
<td>by induction hypothesis</td>
</tr>
<tr>
<td></td>
<td>$R_1(%rdi) = n_1$ and $R_1(%rsp) = R(%rsp)$</td>
</tr>
<tr>
<td></td>
<td>$\forall a \geq R(%rsp), \ M_1(a) = M(a)$</td>
</tr>
<tr>
<td><strong>addq $-8,%rsp$</strong></td>
<td>$R_1', M_1'$</td>
</tr>
<tr>
<td><strong>movq %rdi,(%rsp)</strong></td>
<td>$R_1'= R_1{%rsp \mapsto R(%rsp) - 8}$</td>
</tr>
<tr>
<td></td>
<td>$M_1' = M_1{R(%rsp) - 8 \mapsto n_1}$</td>
</tr>
<tr>
<td><strong>code($e_2$)</strong></td>
<td>$R_2, M_2$</td>
</tr>
<tr>
<td></td>
<td>by induction hypothesis</td>
</tr>
<tr>
<td></td>
<td>$R_2(%rdi) = n_2$ and $R_2(%rsp) = R(%rsp) - 8$</td>
</tr>
<tr>
<td></td>
<td>$\forall a \geq R(%rsp) - 8, \ M_2(a) = M_1'(a)$</td>
</tr>
<tr>
<td><strong>movq (%rsp),%rsi</strong></td>
<td>$R', M_2$</td>
</tr>
<tr>
<td><strong>addq $8,%rsp$</strong></td>
<td>$R'(%rdi) = n_1 + n_2$</td>
</tr>
<tr>
<td><strong>addq %rsi,%rdi</strong></td>
<td>$R'(%rsp) = R(%rsp) - 8 + 8 = R(%rsp)$</td>
</tr>
<tr>
<td></td>
<td>$\forall a \geq R(%rsp)$,</td>
</tr>
<tr>
<td></td>
<td>$M_2(a) = M_1'(a) = M_1(a) = M(a)$</td>
</tr>
</tbody>
</table>
such a proof can be done for a realistic compiler

example: CompCert, an optimizing compiler from C to PowerPC, ARM, RISC-V, and x86, has been formally verified using the Coq proof assistant

see http://compcert.inria.fr/
• lab 2
  • a mini-Python interpreter
  • in Java or OCaml (your choice)
  • take your time to read and understand the code that is provided

• lecture 3
  • parsing

> ./mini-python tests/good/pascal.py
  *
  **
  ***
  ****
  *****
  ******
  *******
  ********
  000000*
  00000**
  00000***
  00000****
  00000*****
  00000******
  00000*******
  00000********
  00000*********
  00000**********