Reachability and error diagnosis in LR(1) parsers

François Pottier

Inria, Paris
February 15, 2016

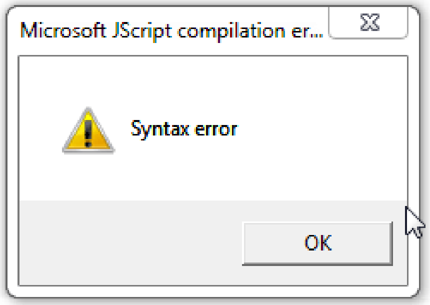
40 years ago

While it is at least honest, a compiler that ***quits*** *upon first detecting an error* ***will not be popular*** *with its users.*

Many runs *may be required just to remove trivial* ***keypunching*** *errors from a program.*

– James J. Horning, "What the compiler should tell the user" (1976)

Today



The finest tools have shortcomings

```
let f x == 3
```

```
$ ocamlc -c f.ml
```

```
File "f.ml", line 1, characters 8-10:
Error: Syntax error
```

Diagnosis using yacc's error token is hard to get right

```
module StringSet = Set.Make(String)
let add (x : int) (xs : StringSet) =
  StringSet.add (string_of_int x) xs
```

```
$ ocamlc -v
The Objective Caml compiler, version 3.10.0
$ ocamlc -c s.ml
```

```
File "s.ml", line 2, characters 33-34:
Syntax error: ')' expected
File "s.ml", line 2, characters 18-19:
This '(' might be unmatched
```

Diagnosis using yacc's error token is hard to get right

```
module StringSet = Set.Make(String)
let add (x : int) (xs : StringSet) =
  StringSet.add (string_of_int x) xs
```

```
$ ocamlc -v
The OCaml compiler, version 4.02.1
$ ocamlc -c s.ml
```

```
File "s.ml", line 2, characters 33-34:
Error: Syntax error: type expected.
```

Should this particular error be so hard to explain ?

```
$ echo "implementation: LET LIDENT LPAREN LIDENT COLON UIDENT RPAREN" \
> | menhir --lalr --interpret-error parser.mly
```

Should this particular error be so hard to explain ?

```
$ echo "implementation: LET LIDENT LPAREN LIDENT COLON UIDENT RPAREN" \
> | menhir --lalr --interpret-error parser.mly
```

```
##
## Ends in an error in state: 194.
##
## mod_ext_longident -> mod_ext_longident . DOT UIDENT [ LPAREN DOT ]
## mod_ext_longident -> mod_ext_longident . LPAREN mod_ext_longident
                                            RPAREN [ LPAREN DOT ]
## type_longident -> mod_ext_longident . DOT LIDENT [ ... 63 tokens ]
```

Only 3 continuations are possible at this point.

The last one is a good suggestion, as it allows a more interesting reduction.

Should this particular error be so hard to explain ?

We might hope to see this :

```
module StringSet = Set.Make(String)
let add (x : int) (xs : StringSet) =
  StringSet.add (string_of_int x) xs
```

```
$ ocamlc -v
The OCaml compiler, version 4.242640687
$ ocamlc -c s.ml
```

```
File "s.ml", line 2, characters 33-34:
Syntax error: ill-formed type.
Up to this point, an extended module path has been recognized.
If this path is complete, then at this point,
a dot '.', followed with a type constructor, is expected.
```

What's the idea ?

Jeffery's idea (2003) :

- Associate a handwritten diagnostic message...
- ...with this invalid sentence (`LET LIDENT LPAREN LIDENT COLON UIDENT RPAREN`).
- Let a tool translate the sentence to a state number (194).

This way, build a collection of state/message pairs...

What's the idea ?

Jeffery's idea (2003) :

- Associate a handwritten diagnostic message...
- ...with this invalid sentence (`LET LIDENT LPAREN LIDENT COLON UIDENT RPAREN`).
- Let a tool translate the sentence to a state number (194).

This way, build a collection of state/message pairs...

...BY HAND.

Oops

Yea, right.

```
$ menhir --lalr -lg 1 -la 1 parser.mly
```

```
Grammar has 206 nonterminal symbols, among which 7 start symbols.
Grammar has 118 terminal symbols.
Grammar has 749 productions.
Built an LR(1) automaton with 1551 states.
```

Research questions

This raises two obvious questions :

- How to come up with a collection of sentences that covers all error states ?
- How to maintain this collection as the grammar evolves ?

Research questions

This raises two obvious questions :

- How to come up with a collection of sentences that covers all error states ?
- How to maintain this collection as the grammar evolves ?

An unanticipated question came up during this work :

- Is it possible to write an accurate diagnostic message for every error state ?

Per-state diagnostic messages ?

Menhir's reachability algorithm and new features

CompCert's new diagnostic messages

Conclusion

How to diagnose syntax errors ?

Jeffery's idea (2003) :

> *Choose a diagnostic message based on the LR automaton's state, ignoring its stack entirely.*

Is this a reasonable idea ?

Let's have a look at a few example situations...

Is this a reasonable idea ? – Yes

Sometimes, yes, clearly the state alone contains enough information.

```
int f (int x) { do {} while (x--) }
```

The error is detected in a state that looks like this :

```
statement: DO statement WHILE LPAREN expr RPAREN . SEMICOLON [...]
```

It is easy enough to give an accurate message :

```
$ ccomp -c dowhile.c
```

```
dowhile.c:1:34: syntax error after ')' and before '}'.
Ill-formed statement.
At this point, a semicolon ';' is expected.
```

Is this a reasonable idea ? – Yes, it seems... ?

Here is another example where things seem to work out as hoped :

```
int f (int x) { return x + 1 }
```

The error is detected in a state that looks like this :

```
expr -> expr . COMMA assignment_expr [ SEMICOLON COMMA ]
expr? -> expr . [ SEMICOLON ]
```

We decide to omit the first possibility, and say a semicolon is expected.

```
$ ccomp -c return.c
```

```
return.c:1:29: syntax error after '1' and before '}'.
Up to this point, an expression has been recognized:
  'x + 1'
If this expression is complete,
then at this point, a semicolon ';' is expected.
```

Yet, ',' and ';' are clearly not the only permitted futures ! What is going on ?

Is this a reasonable idea ? – Uh, oh...

Let us change just the incorrect token in the previous example :

```
int f (int x) { return x + 1 2; }
```

The error is now detected in a different state, which looks like this :

```
postfix_expr -> postfix_expr . LBRACK expr RBRACK [ ... ]
postfix_expr -> postfix_expr . LPAREN arg_expr_list? RPAREN [ ... ]
postfix_expr -> postfix_expr . DOT general_identifier [ ... ]
postfix_expr -> postfix_expr . PTR general_identifier [ ... ]
postfix_expr -> postfix_expr . INC [ ... ]
postfix_expr -> postfix_expr . DEC [ ... ]
unary_expr -> postfix_expr . [ SEMICOLON RPAREN and 34 more tokens ]
```

Based on this information, what diagnostic message can one propose ?

Is this a reasonable idea ? – No !

Based on this, the diagnostic message could say that :

- The “postfix expression” `x + 1` can be continued in 6 different ways ;
- Or maybe this “postfix expression” forms a complete “unary expression”...
- ...and in that case, it could be followed with 36 different tokens...
- among which `';'` appears, but also `')'`, `']'`, `'}'`, and others !

So,

- there is a lot of worthless information,
- yet there is still not enough information :
- we cannot see that `';'` is permitted, while `')'` is not.

The missing information is not encoded in the state : it is buried in the stack.

Two problems

We face two problems :

- depending on which incorrect token we look ahead at, the error is detected in different states ;
- in some of these states, there is not enough information to propose a good diagnostic message.

What can we do about this ?

We propose two solutions to these problems :

- Selective duplication.
  In the grammar, distinguish "expressions that can be followed with a semicolon", "expressions that can be followed with a closing parenthesis", etc.
  (Uses Menhir's expansion of parameterized nonterminal symbols.)

  This fixes the problematic states by building more information into them.
- Reduction on error.
  In the automaton, perform one more reduction to get us out of the problematic state before the error is detected.
  (Uses Menhir's new `%on_error_reduce` directive.)

  This avoids the problematic states.

How do we know what we are doing ?

But :

- how do we find all states where an error can be detected ?
  - in a canonical LR(1) automaton, this is easy...
  - in a non-canonical automaton and in the presence of conflicts, it is not !
- after tweaking the grammar or automaton, how do we know for sure that we have fixed or avoided the problematic states ?

We need tool support.

Per-state diagnostic messages ?

Menhir's reachability algorithm and new features

CompCert's new diagnostic messages

Conclusion

Finding all error states

How do we find all states where an error can be detected ?

- if the grammar is LR(1) and the automaton is canonical,
  then they are exactly the targets of terminal transitions.
- no longer true if the grammar has conflicts or the automaton is noncanonical !

Finding all error states

How do we find all states where an error can be detected ?

- if the grammar is LR(1) and the automaton is canonical, then they are exactly the targets of terminal transitions.
- no longer true if the grammar has conflicts or the automaton is noncanonical !

For every state s and terminal symbol z, if (s, z) is an error entry, we must ask :

- is the configuration (s, z) reachable ?

We need a reachability algorithm.

The algorithm's specification : a big-step semantics of LR(1) automata

Init

$$s \xrightarrow{\epsilon \,/\, \epsilon}\!\!\!\!\rightarrow s\,[z]$$

Step-Terminal

$$\frac{s \xrightarrow{\alpha \,/\, w}\!\!\!\!\rightarrow s'\,[z] \qquad \mathcal{A} \vdash s' \xrightarrow{z} s''}{s \xrightarrow{\alpha z \,/\, wz}\!\!\!\!\rightarrow s''\,[z']}$$

Step-NonTerminal

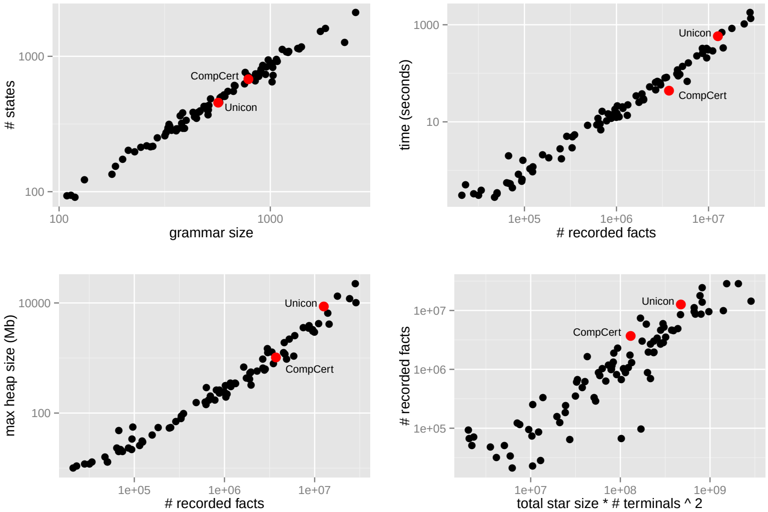
$$\frac{\begin{array}{cc} s \xrightarrow{\alpha \,/\, w}\!\!\!\!\rightarrow s'\,[z] & \mathcal{A} \vdash s' \xrightarrow{A} s'' \\ s' \xrightarrow{A \,/\, w'} s''\,[z'] & z = \mathrm{first}(w'z') \end{array}}{s \xrightarrow{\alpha A \,/\, ww'}\!\!\!\!\rightarrow s''\,[z']}$$

Reduce

$$\frac{\begin{array}{cc} \mathcal{A} \vdash s \xrightarrow{A} s'' & s \xrightarrow{\alpha \,/\, w}\!\!\!\!\rightarrow s'\,[z] \\ \multicolumn{2}{c}{\mathcal{A} \vdash s' \text{ reduces } A \to \alpha \text{ on } z} \end{array}}{s \xrightarrow{A \,/\, w} s''\,[z]}$$

Figure: Inductive characterization of the predicates $s \xrightarrow{\alpha \,/\, w}\!\!\!\!\rightarrow s'\,[z]$ and $s \xrightarrow{A \,/\, w} s'\,[z]$.

The algorithm's performance



Menhir's new features

Menhir can now :

- list all states where an error can be detected,
  together with example sentences that cause these errors.

The grammar author :

- manually constructs a diagnostic message for each error state ;
- adjusts the grammar or automaton to make this task easier.

Menhir :

- updates the list of example sentences and messages as the grammar evolves ;
- checks that this list remains correct, irredundant, and complete.

A few figures

(One version of) CompCert's ISO C99 parser :

- 145 nonterminal symbols, 93 terminal symbols, 365 productions ;
- 677-state LALR(1) automaton ;
- 263 error states found in 43 seconds using 1Gb of memory ;
- 150 distinct hand-written diagnostic messages.

Per-state diagnostic messages ?

Menhir's reachability algorithm and new features

CompCert's new diagnostic messages

Conclusion

Show the past, show (some) futures

```
color->y = (sc.kd * amb->y + il.y + sc.ks * is.y * sc.y;
```

```
$ ccomp -c render.c
```

```
render.c:70:57: syntax error after 'y' and before ';'.
Up to this point, an expression has been recognized:
  'sc.kd * amb->y + il.y + sc.ks * is.y * sc.y'
If this expression is complete,
then at this point, a closing parenthesis ')' is expected.
```

Guidelines :

- Show the past : what has been recently understood.
- Show the future : what is expected next...
- ...but do not show every possible future.

Stay where we are

```
multvec_i[i = multvec_j[i] = 0;
```

```
$ ccomp -c subsumption.c
```

```
subsumption.c:71:34: syntax error after '0' and before ';'.
Ill-formed expression.
Up to this point, an expression has been recognized:
  'i = multvec_j[i] = 0'
If this expression is complete,
then at this point, a closing bracket ']' is expected.
```

Guidelines :

- Show where the problem was detected,
- even if the actual error took place earlier.

Show high-level futures ; show enough futures

```
void f (void) { return; }}
```

```
$ gcc -c braces.c
```

```
braces.c:1: error: expected identifier or ‘(’ before ‘}’ token
```

```
$ clang -c braces.c
```

```
braces.c:1:26: error: expected external declaration
```

```
$ ccomp -c braces.c
```

```
braces.c:1:25: syntax error after '}' and before '}'.
At this point, one of the following is expected:
  a function definition; or
  a declaration; or
  a pragma; or
  the end of the file.
```

Show high-level futures ; show enough futures

Guidelines :

- Do not just say what tokens are allowed next :
- instead, say what high-level constructs are allowed.
- List all permitted futures, if that is reasonable.

Show enough futures

```
int f(void) { int x;) }
```

```
$ gcc -c extra.c
```

```
extra.c: In function ‘f’:
extra.c:1: error: expected statement before ‘)’ token
```

```
$ clang -c extra.c
```

```
extra.c:1:7: error: expected expression
```

```
$ ccomp -c extra.c
```

```
extra.c:1:20: syntax error after ’;’ and before ’)’.
At this point, one of the following is expected:
  a declaration; or
  a statement; or
  a pragma; or
  a closing brace ’}’.
```

Show the goal(s)

```
int main (void) { static const x; }
```

```
$ ccomp -c staticconstlocal.c
```

```
staticconstlocal.c:1:31: syntax error after 'const' and before 'x'.
Ill-formed declaration.
At this point, one of the following is expected:
  a storage class specifier; or
  a type qualifier; or
  a type specifier.
```

Guidelines :

- If possible and useful, show the goal.
- Here, we definitely hope to recognize a “declaration”.

Show the goal(s)

```
static const x;
```

```
$ ccomp -c staticconstglobal.c
```

```
staticconstglobal.c:1:13: syntax error after 'const' and before 'x'.
Ill-formed declaration or function definition.
At this point, one of the following is expected:
  a storage class specifier; or
  a type qualifier; or
  a type specifier.
```

Guidelines :

- Show multiple goals when the choice has not been made yet.
- Here, we hope to recognize a “declaration” or a “function definition”.

Per-state diagnostic messages ?

Menhir's reachability algorithm and new features

CompCert's new diagnostic messages

Conclusion

Contribution

- We equip the Menhir parser generator with tools that help :
  - understand and fine-tune the landscape of syntax errors ;
  - build and maintain a complete collection of diagnostic messages.
- We apply this approach to the CompCert C99 (pre-)parser.

You can do it, too !