

Tree-sitter

@maxbrunsfeld

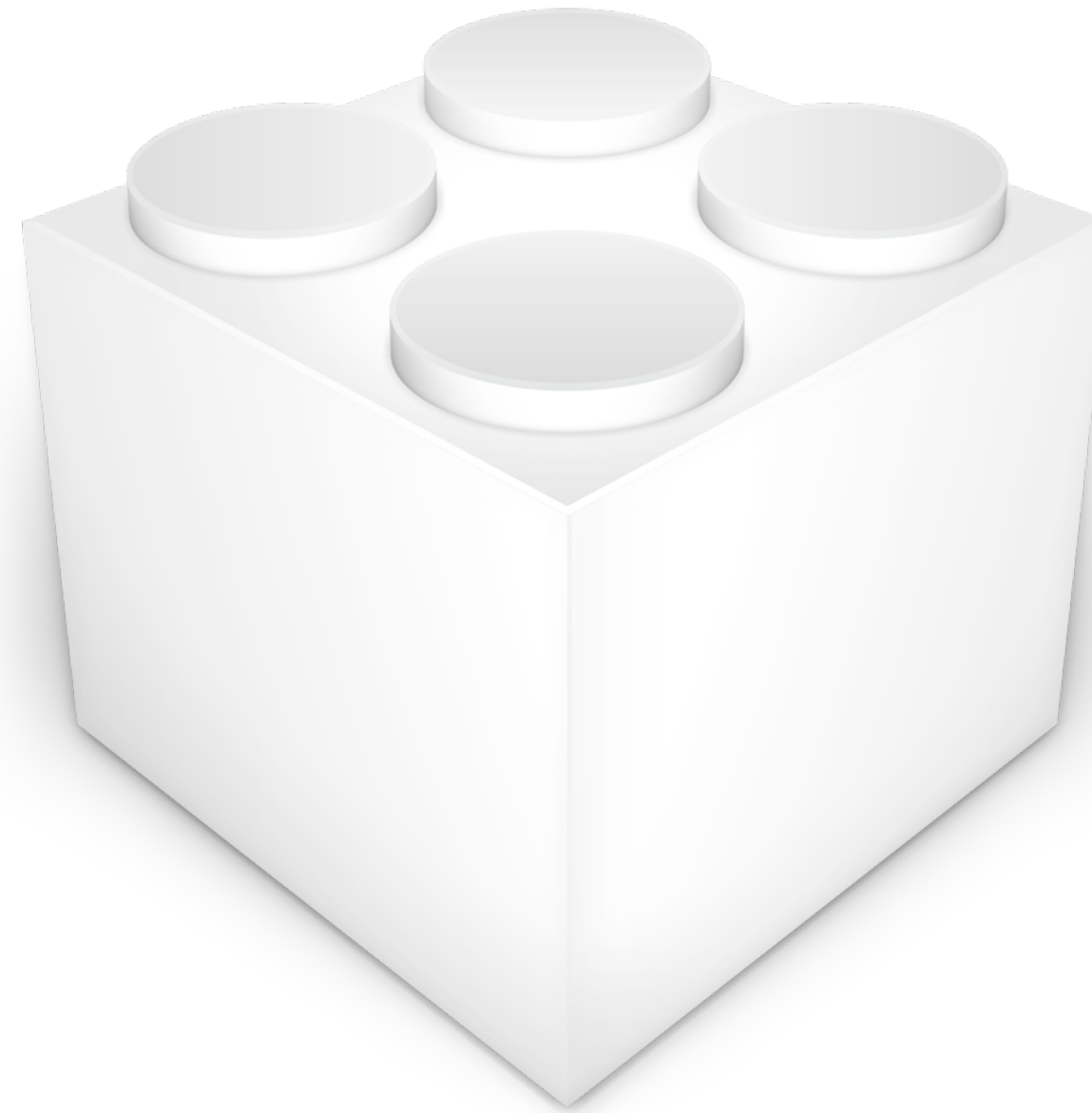
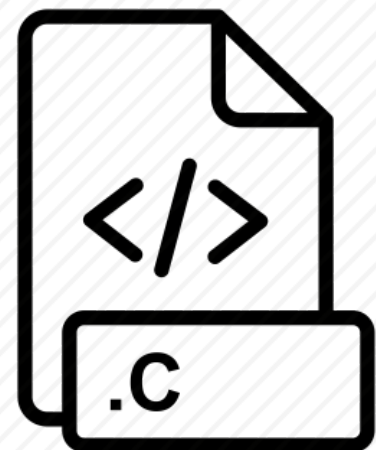
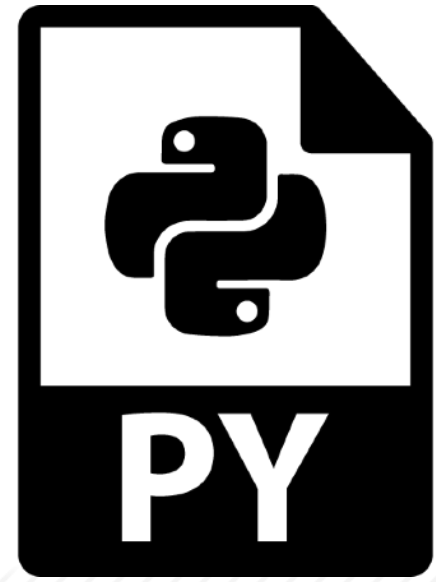


- What is Tree-sitter?
- Why I wrote Tree-sitter
- What we're building with Tree-sitter at GitHub
- Algorithms behind Tree-sitter

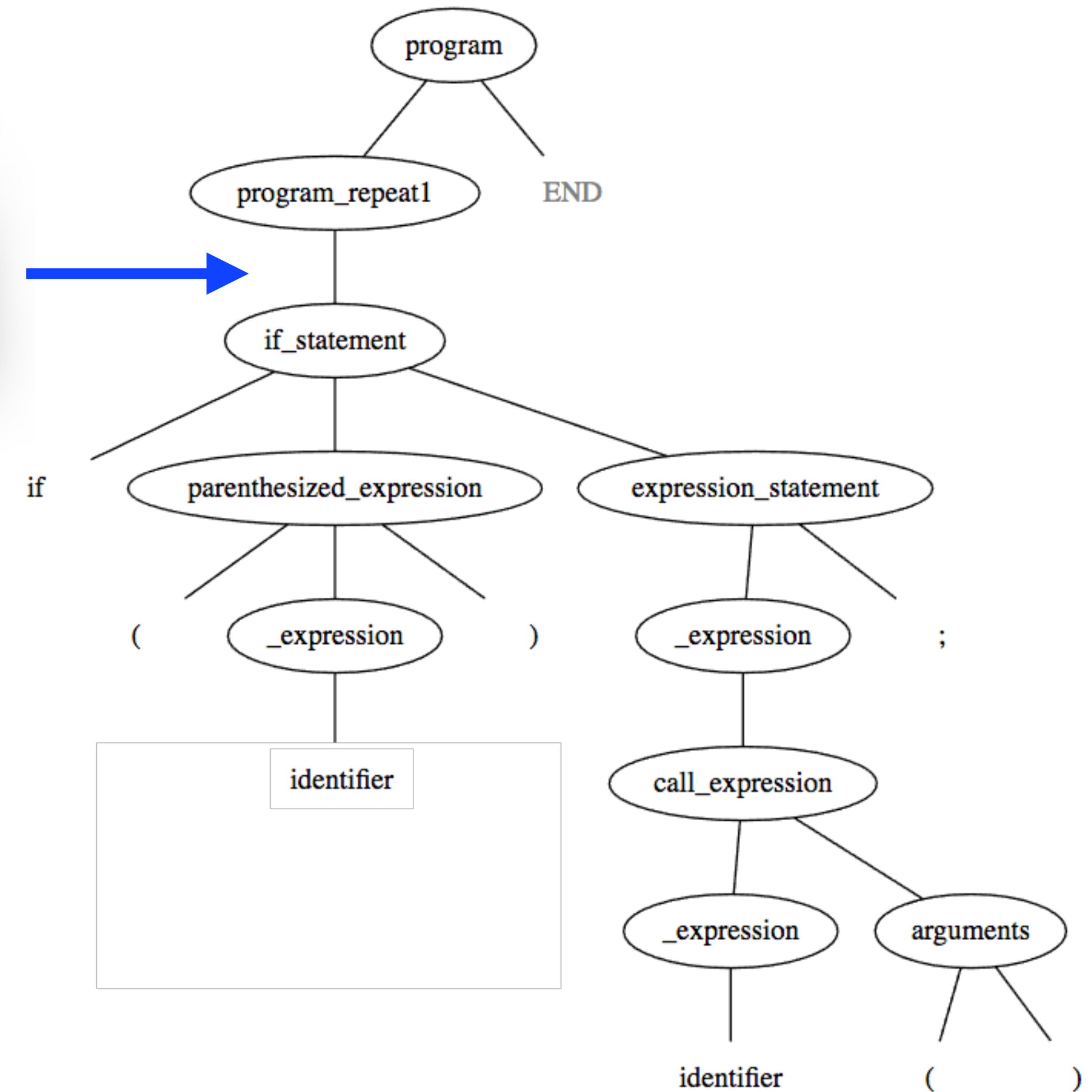


What is Tree-sitter?



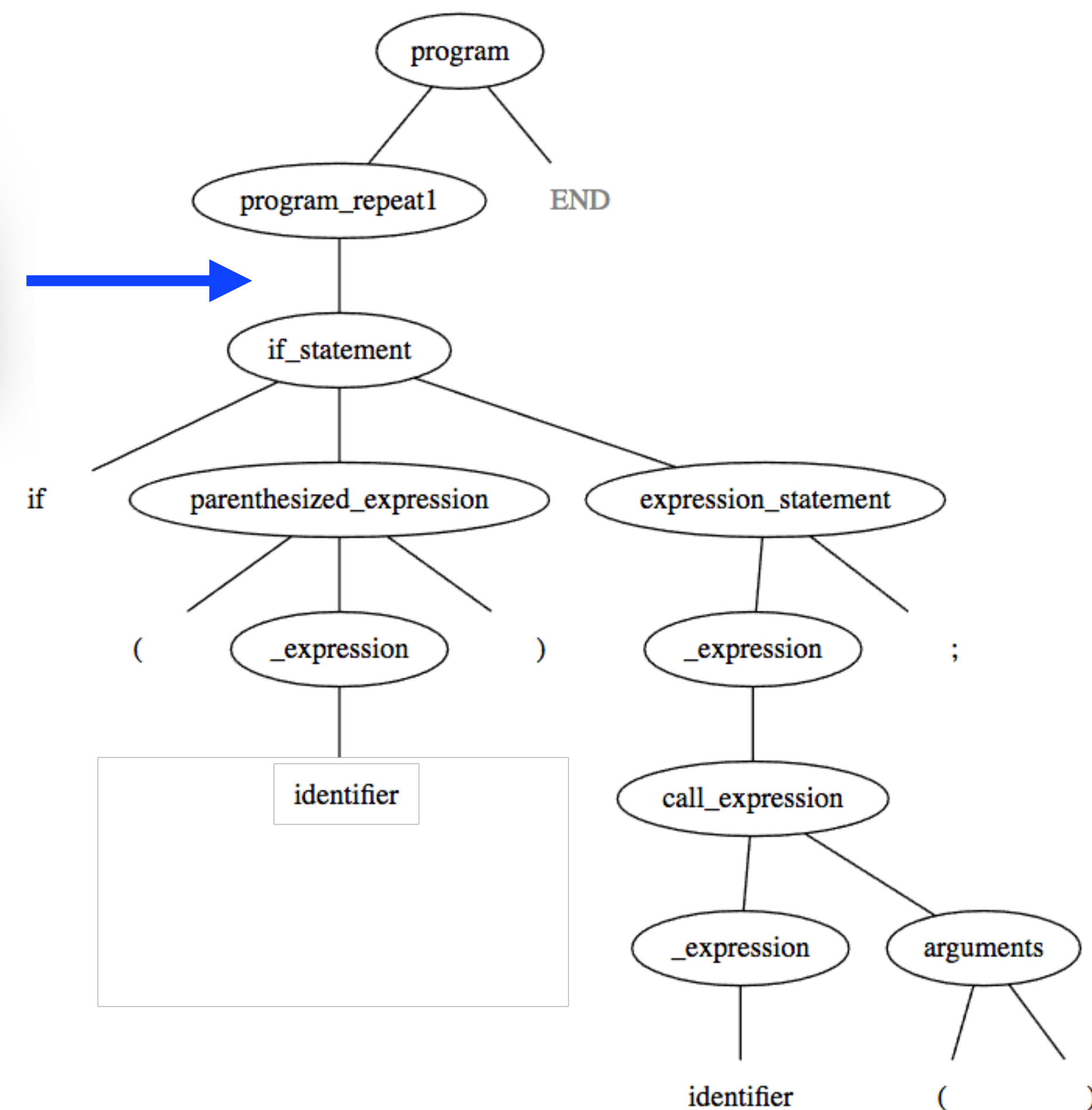
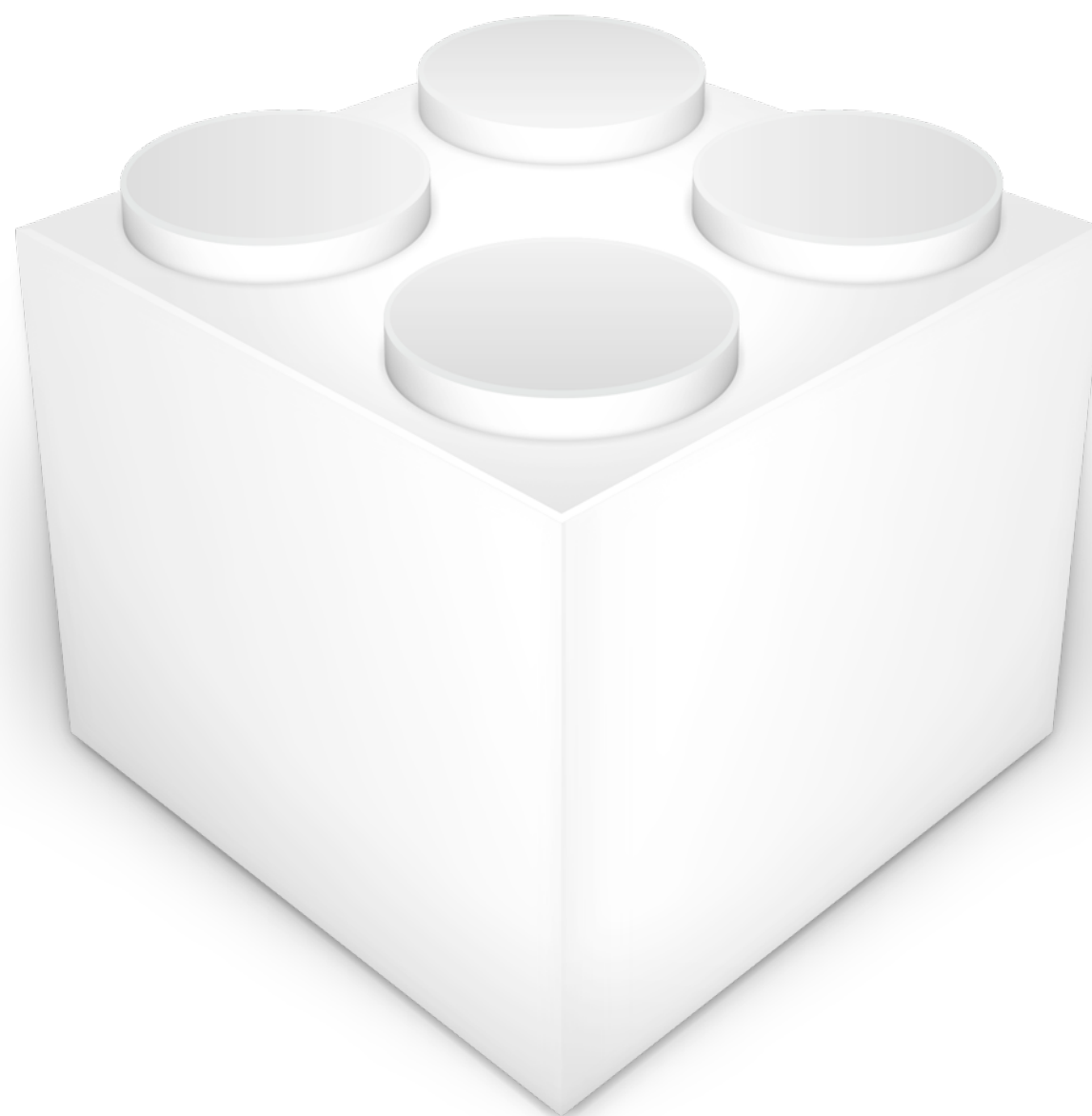
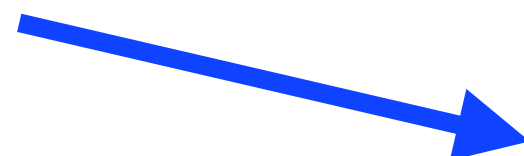


Uniform parsing of
different languages





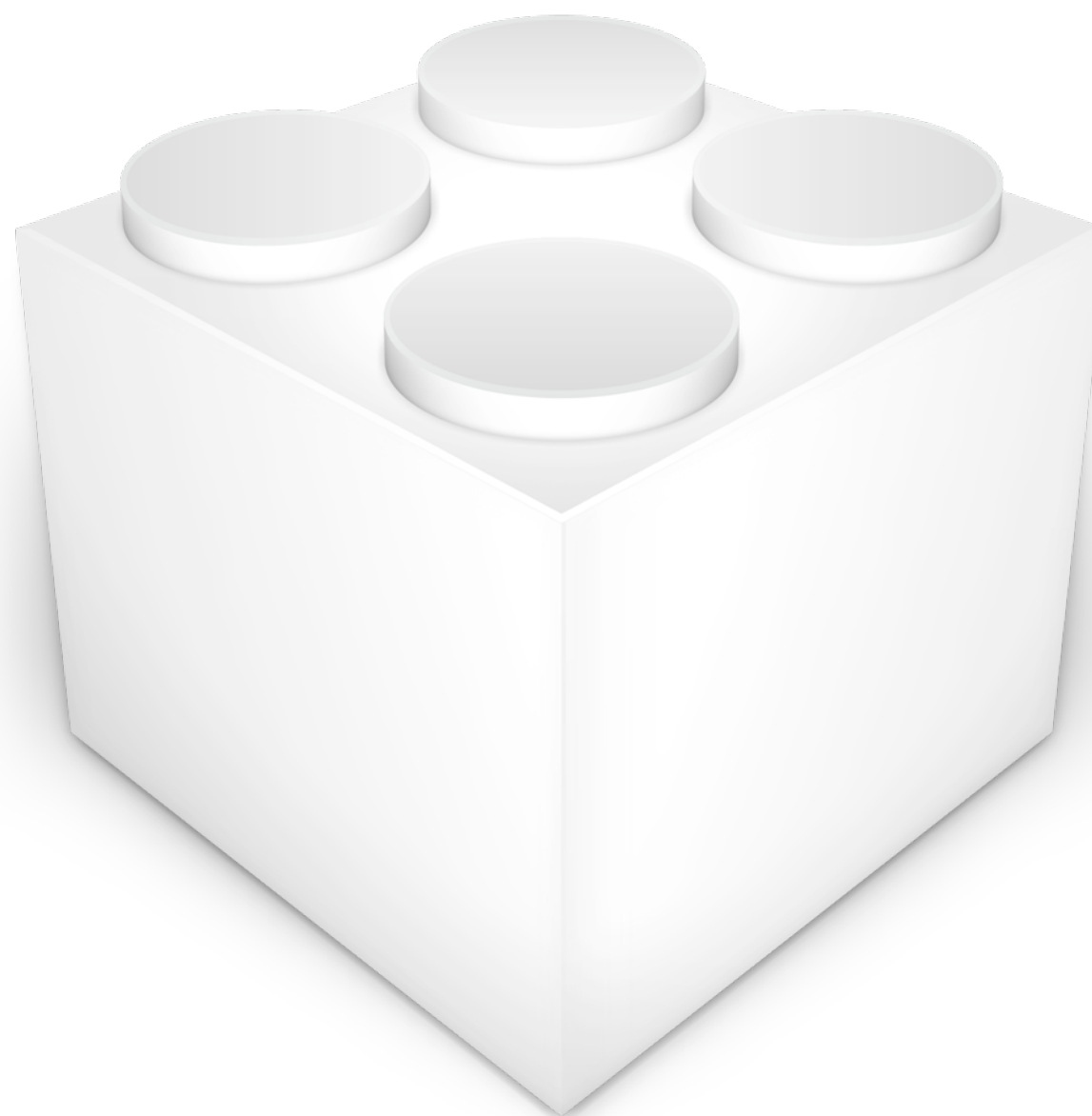
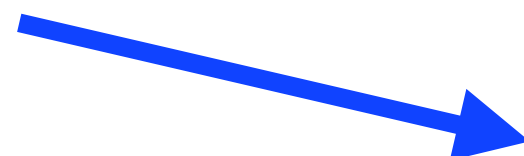
`if (a) c();`



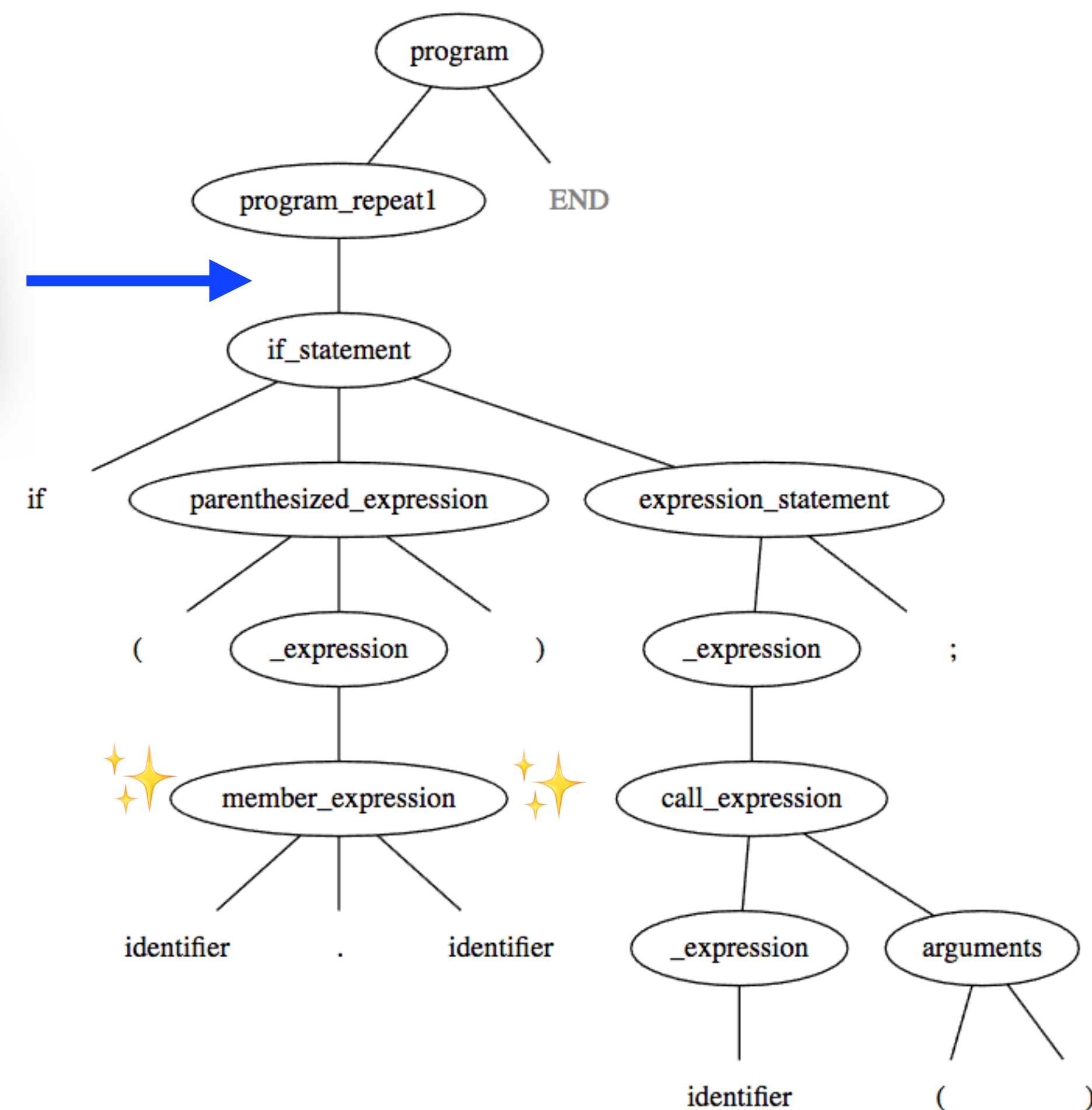
Incremental parsing



`if (a.b) c();`



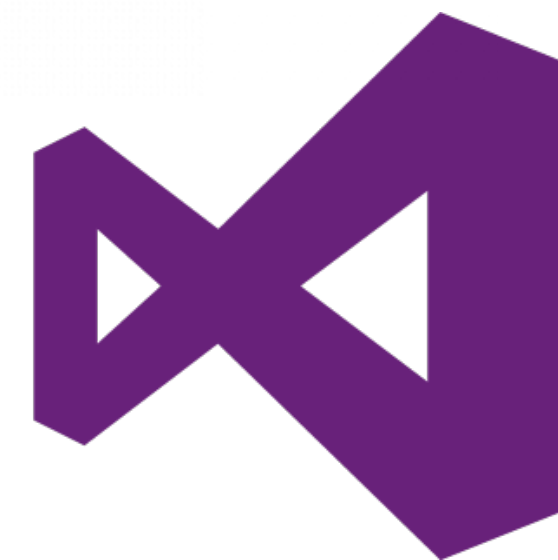
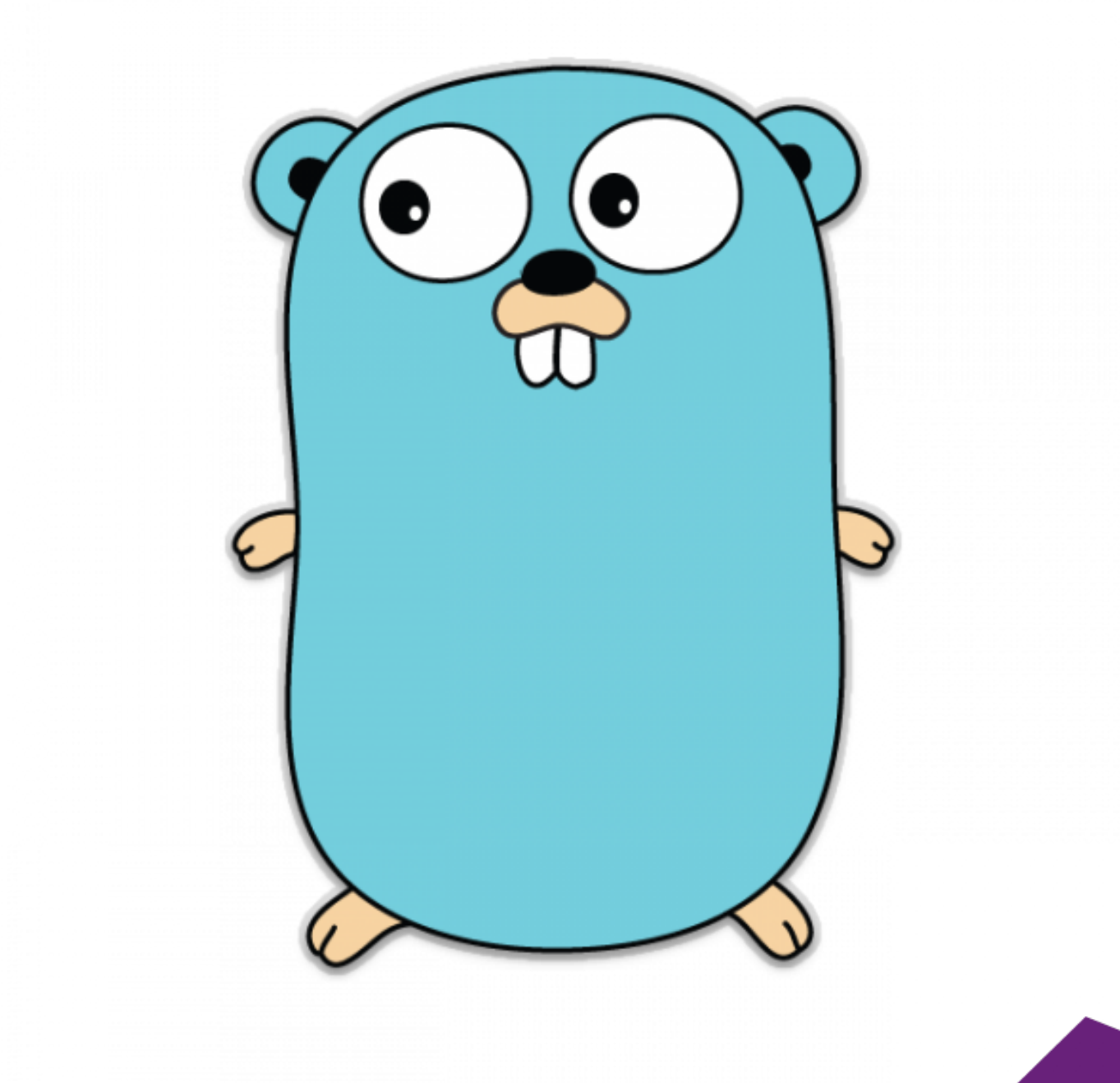
Incremental parsing



Why I wrote Tree-sitter



Single-language code analysis tools



Syntax Highlighting Today

```
3  type Person struct {
4      name string
5      mom *Person
6  }
7
8  func NewPerson(name string, mom *Person) Person {
9      return Person{name: name, mom: mom}
10 }
11
12 func (self *Person) GetName() string {
13     return self.name
14 }
15
16 func (self *Person) GetMom() *Person {
17     return self.mom
18 }
```



Syntax Highlighting Today

```
3  type Person struct {
4      name string
5      mom *Person
6  }
7
8  func NewPerson(name string, mom *Person) Person {
9      return Person{name: name, mom: mom}
10 }
11
12 func (self *Person) GetName() string {
13     return self.name
14 }
15
16 func (self *Person) GetMom() *Person {
17     return self.mom
18 }
```

Why three different colors for types?



Syntax Highlighting Today

```
3 type Person struct {
4     name string
5     mom *Person
6 }
7
8 func NewPerson(name string, mom *Person) Person {
9     return Person{name: name, mom: mom}
10 }
11
12 func (self *Person) GetName() string {
13     return self.name
14 }
15
16 func (self *Person) GetMom() *Person {
17     return self.mom
18 }
```

Why three different colors for types?



Syntax Highlighting Today

```
3  type Person struct {
4      name string
5      mom *Person
6  }
7
8  func NewPerson(name string, mom *Person) Person {
9      return Person{name: name, mom: mom}
10 }
11
12 func (self *Person) GetName() string {
13     return self.name
14 }
15
16 func (self *Person) GetMom() *Person {
17     return self.mom
18 }
```

Why three different colors for types?



Why not use standard parsers?

- Most are too slow for text editors
- Many require extra information that's not available
- Each has its own dependencies



Why use Tree-sitter



- It's fast
- It requires no extra information
- Parsers have no dependencies



What we're doing with Tree-sitter



Syntax Highlighting - Go

```
3  type Person struct {
4      |      name string
5      |      mom *Person
6  }
7
8  func NewPerson(name string, mom *Person) Person {
9      |      return Person{name: name, mom: mom}
10 }
11
12 func (self *Person) GetName() string {
13     |     return self.name
14 }
15
16 func (self *Person) GetMom() *Person {
17     |     return self.mom
18 }
```



Syntax Highlighting - C

```
1  typedef struct {
2      |  char *name;
3      |  Person *mom;
4  } Person;
5
6  Person new_person(char *name, Person *mom) {
7      |  return (Person){.name = name, .mom = mom};
8  }
9
10 const char *get_name(Person *self) {
11     |  return self->name;
12 }
13
14 const Person *get_mom(Person *self) {
15     |  return self->mom;
16 }
```



Syntax Highlighting - C++

```
1  class Person {
2      const std::string name;
3      const Person *mom;
4
5  public:
6      Person(std::string name, const Person *mom)
7          : name(name), mom(mom) {}
8
9      const std::string &get_name() {
10         return this->name;
11     }
12
13     const Person *get_mom() {
14         return this->mom;
15     }
16 };
```



Syntax Highlighting - TypeScript

```
1  class Person {  
2      name: string  
3      mom: Person  
4  
5      constructor(name: string, mom: Person) {  
6          this.name = name;  
7          this.mom = mom;  
8      }  
9  
10     getName() : string {  
11         return this.name;  
12     }  
13  
14     getMom() : Person {  
15         return this.mom;  
16     }  
17 };
```



Syntax Highlighting - Python

```
1  class Person:
2      def __init__(self, name: string, mom: Person):
3          self.name = name
4          self.mom = mom
5
6      def get_name(self):
7          return self.name
8
9      def get_mom(self):
10         return self.mom
11
```



Syntax Highlighting - Ruby

```
1  class Person
2      def initialize(name, mom)
3          @name = name
4          @mom = mom
5      end
6
7      def name
8          @name
9      end
10
11     def mom
12         @mom
13     end
14 end
```



Syntax Highlighting - Rust

```
1  struct Person<'a> {  
2      name: String,  
3      mom: &'a Person<'a>,  
4  }  
5  
6  impl<'a> Person<'a> {  
7      pub fn get_name(&self) -> &String {  
8          &self.name  
9      }  
10  
11     pub fn get_mom(&self) -> &'a Person {  
12         self.mom  
13     }  
14 }
```




```
1  /*! jQuery v3.1.1 | (c) jQuery Foundation | jquery.org/license */
2  !function(a,b){"use strict";"object"==typeof module&&"object"==typeof module.exports?module.exports=a.document?b(a,!0):fu
•   c=b.createElement("script");c.text=a,b.head.appendChild(c).parentNode.removeChild(c)}var q="3.1.1",r=function(a,b){retu
•   b.prevObject=this,b},each:function(a){return r.each(this,a)},map:function(a){return this.pushStack(r.map(this,function(
•   this.prevObject||this.constructor()),push:h,sort:c.sort,splice:c.splice},r.extend=r.fn.extend=function(){var a,b,c,d,e,
•   0!==d&&(g[b]=d));return g},r.extend({expando:"jQuery"+(q+Math.random()).replace(/\D/g,""),isReady:!0,error:function(a){
•   Object]"!==k.call(a))&&(!(b=e(a))|| (c=l.call(b,"constructor")&&b.constructor,"function"==typeof c&&m.call(c)===n))),isE
•   c,d=0;if(w(a)){for(c=a.length;d<c;d++)if(b.call(a[d],d,a[d])===!1)break}else for(d in a)if(b.call(a[d],d,a[d])===!1)bre
•   d,e=[],f=0,g=a.length,h=!c;f<g;f++)d=!b(a[f],f),d!==h&&e.push(a[f]);return e},map:function(a,b,c){var d,e,f=0,h=[];if(w
•   a.apply(b||this,d.concat(f.call(arguments)))},e.guid=a.guid=a.guid||r.guid++,e},now:Date.now,support:o)),"function"==ty
•   b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u="sizzle"+1*new Date,v=a.document,w=0,x=0,y=ha(),z=ha(),A=ha(),B=function(a,b){r
•   c;return-1},J="checked|selected|async|autofocus|autoplay|controls|defer|disabled|hidden|ismap|loop|multiple|open|reador
•   RegExp("^"+K+"+|((?:^|^[^\\]\\])?:\\\\.)*"+K+"+$","g"),Q=new RegExp("^"+K+"*","g"),R=new RegExp("^"+K+"*([>~]|"+K+"
•   RegExp("^:(only|first|last|nth|nth-last)-(child|of-type)(?:\\ "("+K+"*(even|odd|(([+-]|)(\\d*)n|)" +K+"*(?:([+-]|)" +K+"*(\\
•   RegExp("\\\\([\\da-f]{1,6})"+K+"?|("+K+"|\\.)" , "ig"),aa=function(a,b,c){var d="0x"+b-65536;return d!==d||c?b:d<0?String.f
•   a)},{dir:"parentNode",next:"legend"});try{G.apply(D=H.call(v.childNodes),v.childNodes),D[v.childNodes.length].nodeType}
•   d;if(!e&&(b?b.ownerDocument||b:v) !==n&&m(b),b=b||n,p)){if(11!==w&&(l=Z.exec(a)))if(f=l[1]){if(9===w){if(!(j=b.getEleme
•   if("object"!==b.nodeName.toLowerCase()){(k=b.getAttribute("id"))?k=k.replace(ba,ca):b.setAttribute("id",k=u),o=g(a),h=c
•   a[u]=!0,a}function ja(a){var b=n.createElement("fieldset");try{return!!a(b)}catch(c){return!1}finally{b.parentNode&&b.p
•   c=b.nodeName.toLowerCase();return
```



```

1  /*! jQuery v3.1.1 | (c) jQuery Foundation | jquery.org/license */
2  !function(a,b){["use strict";"object"==typeof module&&"object"==typeof module.exports?module.exports=a.document?b(a,!0):fu
   c=b.createElement("script");c.text=a,b.head.appendChild(c).parentNode.removeChild(c)}var q="3.1.1",r=function(a,b){retu
   b.prevObject=this,b},each:function(a){return r.each(this,a)},map:function(a){return this.pushStack(r.map(this,function(
   this.prevObject||this.constructor()),push:h,sort:c.sort,splice:c.splice),r.extend=r.fn.extend=function(){var a,b,c,d,e,
   0!==d&&(g[b]=d));return g},r.extend({expando:"jQuery"+(q+Math.random()).replace(/\D/g,""),isReady:!0,error:function(a){
   Object]"!:=k.call(a))&&(! (b=e(a)) || (c=l.call(b,"constructor")&&b.constructor,"function"==typeof c&&m.call(c)==n)),isE
   c,d=0;if(w(a)){for(c=a.length;d<c;d++)if(b.call(a[d],d,a[d])===!1)break}else for(d in a)if(b.call(a[d],d,a[d])===!1)bre
   d,e=[],f=0,g=a.length,h=!c;f<g;f++)d=!b(a[f],f),d!==h&&.push(a[f]);return e},map:function(a,b,c){var d,e,f=0,h=[];if(w
   a.apply(b||this,d.concat(f.call(arguments))),e.guid=a.guid=a.guid||r.guid++,e,now>Date.now,support:o}),"function"==ty
   b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u="sizzle"+1*new Date,v=a.document,w=0,x=0,y=ha(),z=ha(),A=ha(),B=function(a,b){r
   c;return-1},J="checked|selected|async|autofocus|autoplay|controls|defer|disabled|hidden|ismap|loop|multiple|open|reador
   RegExp("^"+K+"+(?:^|([^\]\])?(?:\\.\.)*"+K+"+$","g"),Q=new RegExp("^"+K+"*","+K"+"*"),R=new RegExp("^"+K+"*([>~]|"+K+"
   RegExp("^:(only|first|last|nth|nth-last)-(child|of-type)(?:\\("+K+"*(even|odd|(([+-]|)(\\d*)n|)" +K+"*(?:([+-]|)" +K+"*(\\
   RegExp("\\\\([\\da-f]{1,6})"+K+"?|(\"+K+\")|.)","ig"),aa=function(a,b,c){var d="0x"+b-65536;return d!=d||c?b:d<0?String.f
   a)},{dir:"parentNode",next:"legend"});try{G.apply(D=H.call(v.childNodes),v.childNodes),D[v.childNodes.length].nodeType}
   d;if(!e&&(b?v.ownerDocument||b:v)!=n&&m(b),b=b||n,p)){if(11!==w&&(l=Z.exec(a)))if(f=l[1]){if(9===w){if(!(j=b.getEleme
   if("object"!==b.nodeName.toLowerCase()){(k=b.getAttribute("id"))?k=k.replace(ba,ca):b.setAttribute("id",k=u),o=g(a),h=c
   a[u]=!0,a}function ja(a){var b=n.createElement("fieldset");try{return!!a(b)}catch(c){return!1}finally{b.parentNode&&b.p
   c=b.nodeName.toLowerCase();return"input"===c&&b.type===a}}function na(a){return function(b){var c=b.nodeName.toLowerCase
   e,f=a([],c.length,b),g=f.length;while(g--)c[e=f[g]]&&(c[e]=!(d[e]=c[e]))}}function qa(a){return a&&"undefined"!==type
   g!==n&&9===g.nodeType&&g.documentElement?(n=g,o=n.documentElement,p=!f(n),v!==n&&(e=n.defaultView)&&e.top!==e&&(e.addEv
   a.appendChild(n.createComment(""))),!a.getElementsByTagName("*").length},c.getElementsByClassName=Y.test(n.getElementsB
   b=a.replace(_,aa);return function(a){var c="undefined"!==typeof a.getAttributeNode&&a.getAttributeNode("id");return c&&c
   b.getElementsByTagName?b.getElementsByTagName(a):c.qsa?b.querySelectorAll(a):void 0}:function(a,b){var c,d=[],e=0,f=b.g
   msallowcapture=' '><option selected=' '></option></select>",a.querySelectorAll("[msallowcapture^=' '']").length&&q.push("[*
   disabled='disabled'><option/></select>";var
   b=n.createElement("input");b.setAttribute("type","hidden"),a.appendChild(b).setAttribute("name","D"),a.querySelectorAll
   ||o.msMatchesSelector))&&j(function(a){c.disconnectedMatch=s.call(a,"*"),s.call(a,"[s!='']:x"),r.push("!=",N)}),q=q.le

```

Syntax Highlighting - Performance

```
$ wc -l react.dev.js
```

```
20599 react.dev.js
```

```
$ tree-sitter parse react.dev.js --time
```

```
react.dev.js 69ms
```



Code Folding

```
1 > typedef struct {...} Person;  
5  
6 > Person new_person(char *name, Person *mom) {...}  
9  
10 > const char *get_name(Person *self) {...}  
13  
14 > const Person *get_mom(Person *self) {...}  
17
```



Code Folding

```
1  typedef struct {
2      | char *name;
3      | Person *mom;
4  } Person;
5
6  Person new_person(char *name,
7      | | | | | | | | | | Person *mom) {
8      | return (Person){.name = name, .mom = mom};
9  }
10
11 const char *get_name(Person *self) {
12     | return self->name;
13 }
14
15 const Person *get_mom(Person *self) {
16     | return self->mom;
17 }
```

Indentation doesn't always
match syntax



Code Folding

```
1  typedef struct {
2      |  char *name;
3      |  Person *mom;
4  } Person;
5
6  Person new_person(char *name, Person *mom) {
7      |  return (Person){.name = name, .mom = mom};
8  }
9
10 const char *get_name(Person *self) {
11     |  return self->name;
12 }
13
14 const Person *get_mom(Person *self) {
15     |  return self->mom;
16 }
```



Extend Selection

```
4  it('reports a tag boundary at relevant nodes in the tree', function () {
5    assert.deepEqual(getTokens(buffer, iterator), [
6      [
7        {
8          text: ' ',
9          scopes: [
10           {
11             value: 'source.js'
12           }
13         ]
14       },
15       {
16         text: 'function',
17         scopes: [
18           {
19             value: 'source.js'
20           },
21           {
22             value: 'meta.function.js'
23           },
24           {
25             value: 'storage.type.function.js'
26           }
27         ]
28       },
29       {
30         text: ' ',
31         scopes: [
32           {
33             value: 'source.js'
34           },
35           {
36             value: 'meta.function.js'
37           }
38         ]
39       }
40     ]
41   })
42 })
```



Pull Request Table of Contents

Jump to files or symbols

Filter changed files or symbols

command_pre_push.go commands/command_pre_push.go +2 -4	
prePushCommand function <td></td>	
command_push.go commands/command_push.go +2 -31	
uploadsBetweenRefAndRemote function <td></td>	
uploadLeftOrAll function <td></td>	
uploader.go commands/uploader.go +79 -8	
uploadLeftOrAll function <td></td>	
(*gitScannerLockables) Contains method <td></td>	
newUploadContext function <td></td>	
(*uploadContext) scannerError method <td></td>	



How Tree-sitter Works



Writing a Grammar

grammar.js

```
414 if_statement: $ => seq(  
415   'if',  
416   optional(seq($_.simple_statement, ';')),  
417   $_.expression,  
418   $.block,  
419   optional(seq(  
420     'else',  
421     choice($.block, $.if_statement)  
422   ))  
423 ),  
424  
425 for_statement: $ => seq(  
426   'for',  
427   optional(choice($_.expression, $.for_clause, $.range_clause)),  
428   $.block  
429 ),  
430  
431 for_clause: $ => seq(  
432   optional($_.simple_statement),  
433   ';',  
434   optional($_.expression),  
435   ';',  
436   optional($_.simple_statement)  
437 ),  
438  
439 range_clause: $ => seq(  
440   optional(seq(  
441     $.expression_list,  
442     choice('=', ':=')  
443   )),  
444   'range',  
445   $_.expression  
446 ),
```



Writing a Grammar

parser.c

```
1535 case 24:
1536     if (('0' <= lookahead && lookahead <= '9') ||
1537         ('A' <= lookahead && lookahead <= 'F') ||
1538         ('a' <= lookahead && lookahead <= 'f'))
1539         ADVANCE(13);
1540     LEX_ERROR();
1541 case 25:
1542     ACCEPT_TOKEN(anon_sym_LPAREN);
1543 case 26:
1544     ACCEPT_TOKEN(anon_sym_RPAREN);
1545 case 27:
1546     if (lookahead == '=')
1547         ADVANCE(8);
1548     ACCEPT_TOKEN(anon_sym_STAR);
1549 case 28:
1550     if (lookahead == '+')
1551         ADVANCE(29);
1552     if (lookahead == '=')
1553         ADVANCE(8);
1554     ACCEPT_TOKEN(anon_sym_PLUS);
1555 case 29:
1556     ACCEPT_TOKEN(anon_sym_PLUS_PLUS);
1557 case 30:
1558     ACCEPT_TOKEN(anon_sym_COMMA);
```



Writing a Grammar

parser.c

```
13800 [2] = {
13801     [sym_identifier] = ACTIONS(109),
13802     [sym_comment] = ACTIONS(105),
13803 },
13804 [3] = {
13805     [ts_builtin_sym_end] = ACTIONS(111),
13806     [sym_comment] = ACTIONS(105),
13807 },
13808 [4] = {
13809     [sym_import_declaration] = STATE(10),
13810     [sym__top_level_declaration] = STATE(11),
13811     [sym__declaration] = STATE(12),
13812     [sym_const_declaration] = STATE(13),
13813     [sym_var_declaration] = STATE(13),
13814     [sym_function_declaration] = STATE(12),
13815     [sym_method_declaration] = STATE(12),
13816     [sym_type_declaration] = STATE(13),
13817     [aux_sym_source_file_repeat1] = STATE(14),
13818     [aux_sym_source_file_repeat2] = STATE(15),
13819     [ts_builtin_sym_end] = ACTIONS(113),
13820     [anon_sym_import] = ACTIONS(115),
13821     [anon_sym_const] = ACTIONS(117),
13822     [anon_sym_var] = ACTIONS(119),
13823     [anon_sym_func] = ACTIONS(121),
13824     [anon_sym_type] = ACTIONS(123),
13825     [sym_comment] = ACTIONS(105),
13826 },
13827 [5] = {
13828     [sym_import_spec] = STATE(1238),
13829     [sym__string_literal] = STATE(1239),
13830     [anon_sym_LPAREN] = ACTIONS(125),
13831     [anon_sym_DOT] = ACTIONS(127),
13832     [sym_identifier] = ACTIONS(129),
13833     [sym_raw_string_literal] = ACTIONS(131),
13834     [sym_interpreted_string_literal] = ACTIONS(131),
13835     [sym_comment] = ACTIONS(105),
13836 },
```



Using the parser

```
1  #include "tree_sitter/runtime.h"
2
3  ✓ int main() {
4      const TSDocument *document = ts_document_new();
5      const TSLanguage *javascript_language = tree_sitter_javascript();
6
7      ts_document_set_language(document, javascript_language);
8      ts_document_set_input_string(document, "var x = 1; print(x);");
9      ts_document_parse(document);
10
11     TSNode root_node = ts_document_root_node(document);
12     assert(ts_node_child_count(root_node) == 2);
13 }
```



Algorithms



Existing Research

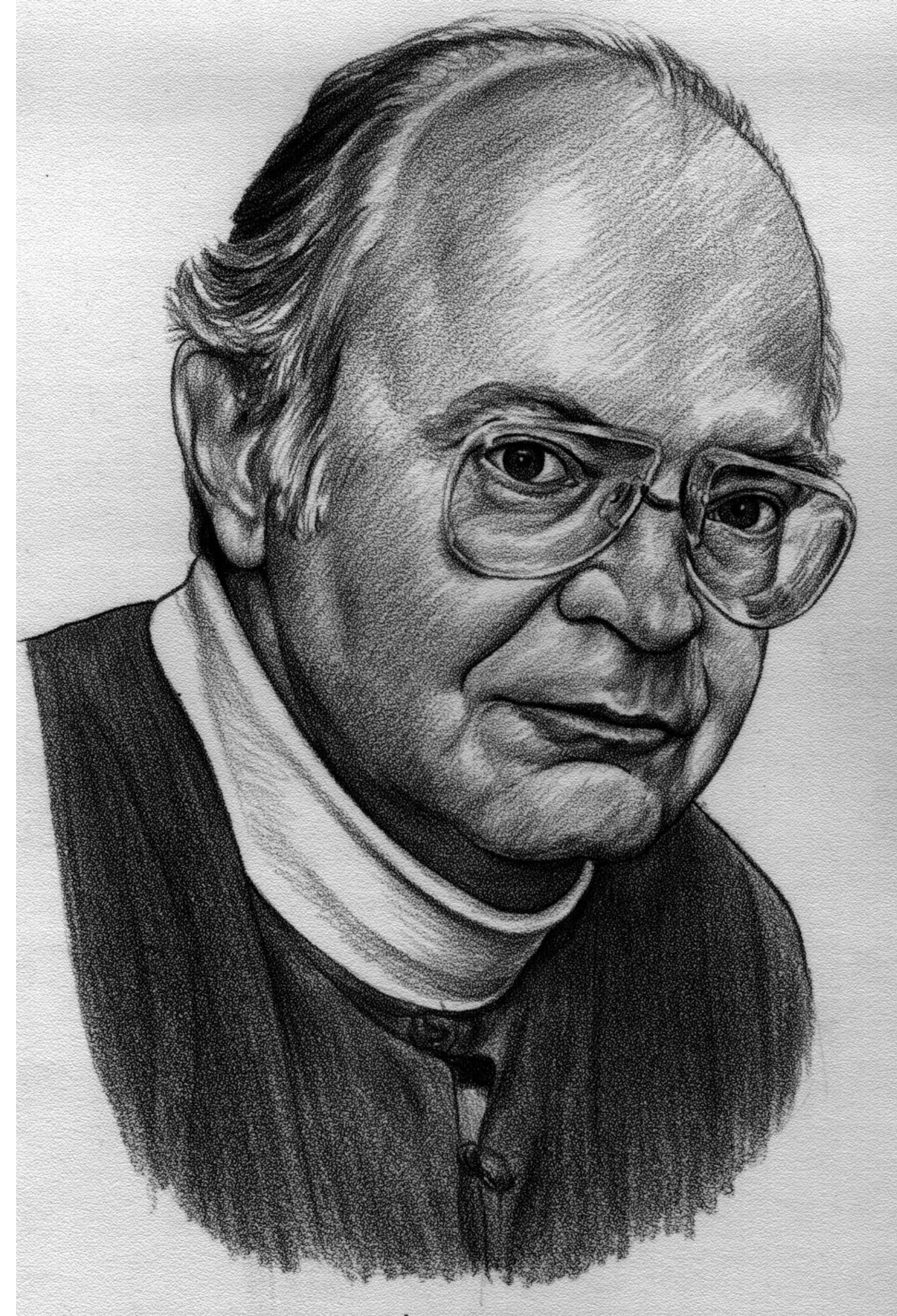


*Practical Algorithms for Incremental Software Development
Environments*

Tim A. Wagner 1998



LR Parsing



LR Parsing

Text: $x * y + z$



LR Parsing

Text: $x * y + z$
 ↑

Stack:

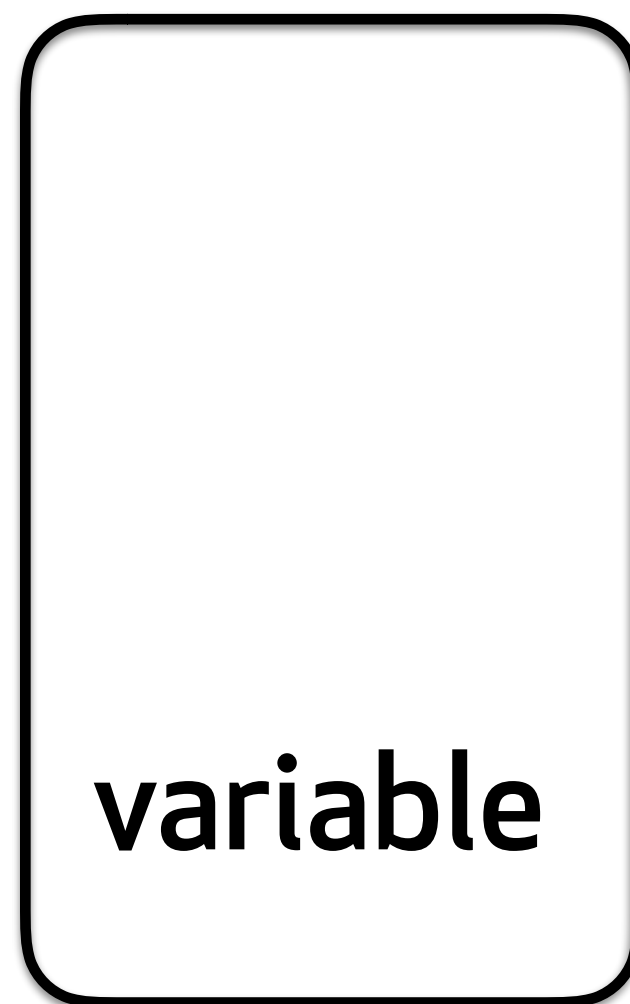


LR Parsing

Text: x * y + z



Stack:

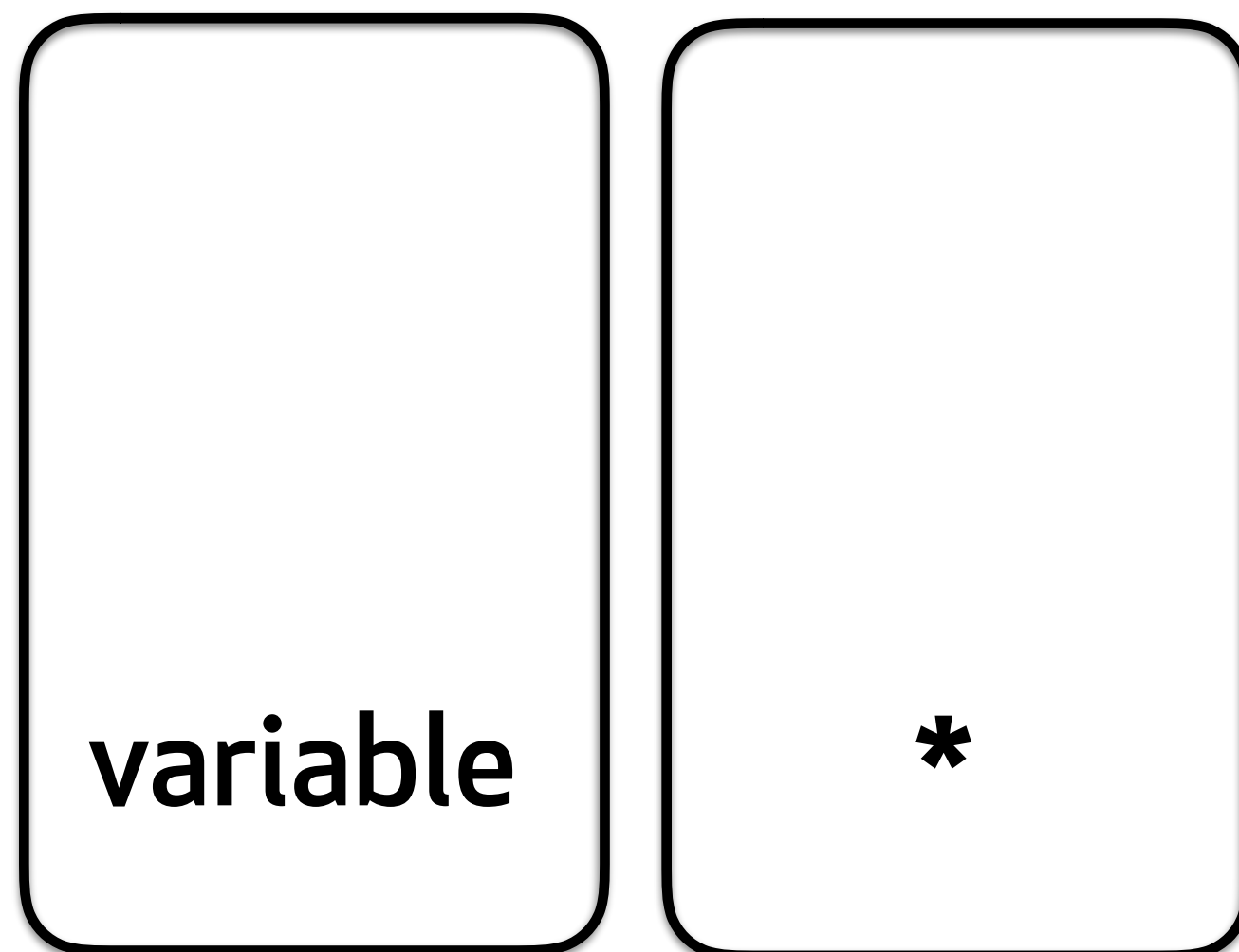


LR Parsing

Text: x * y + z



Stack:

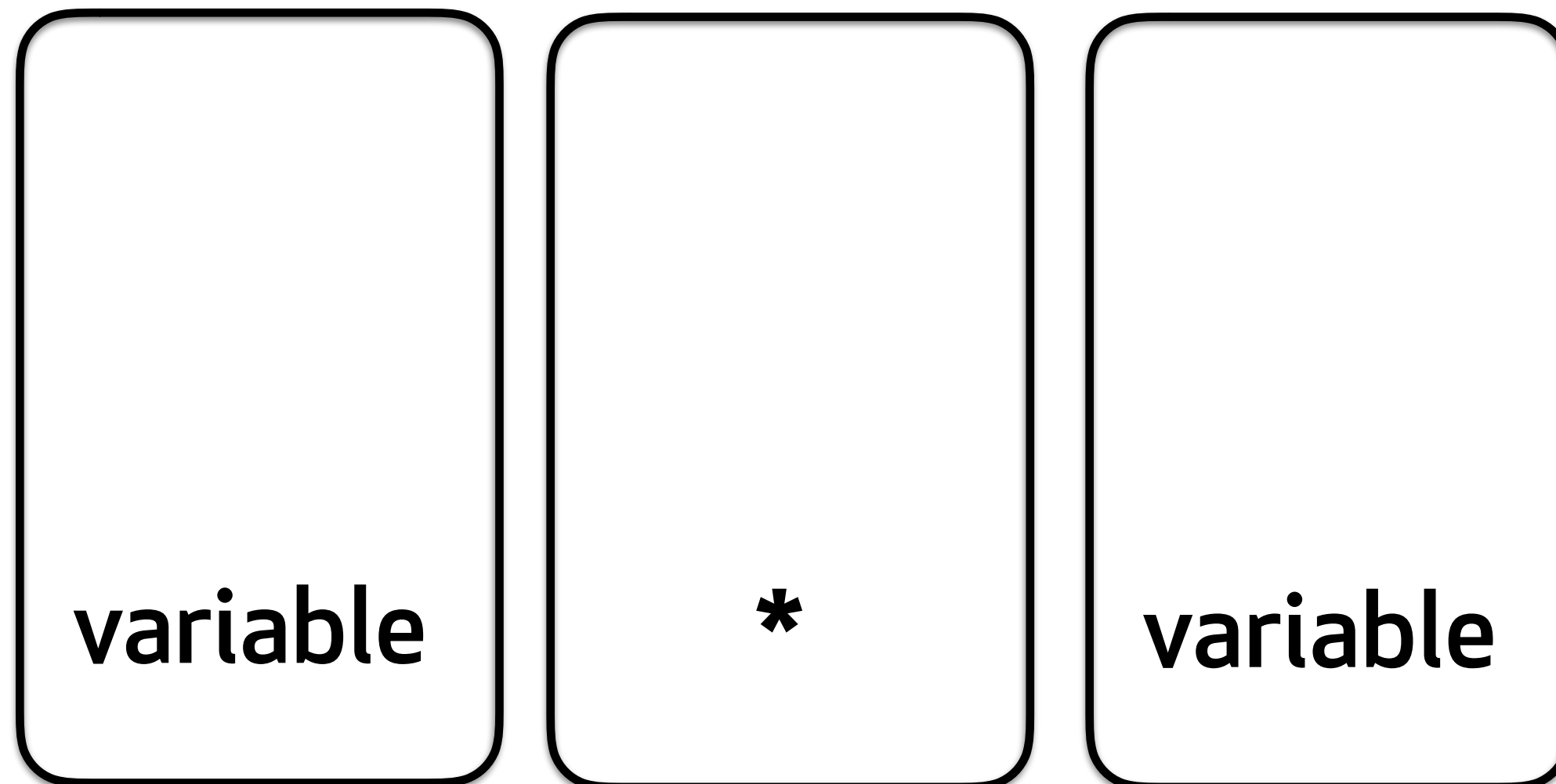


LR Parsing

Text: x * y + z



Stack:

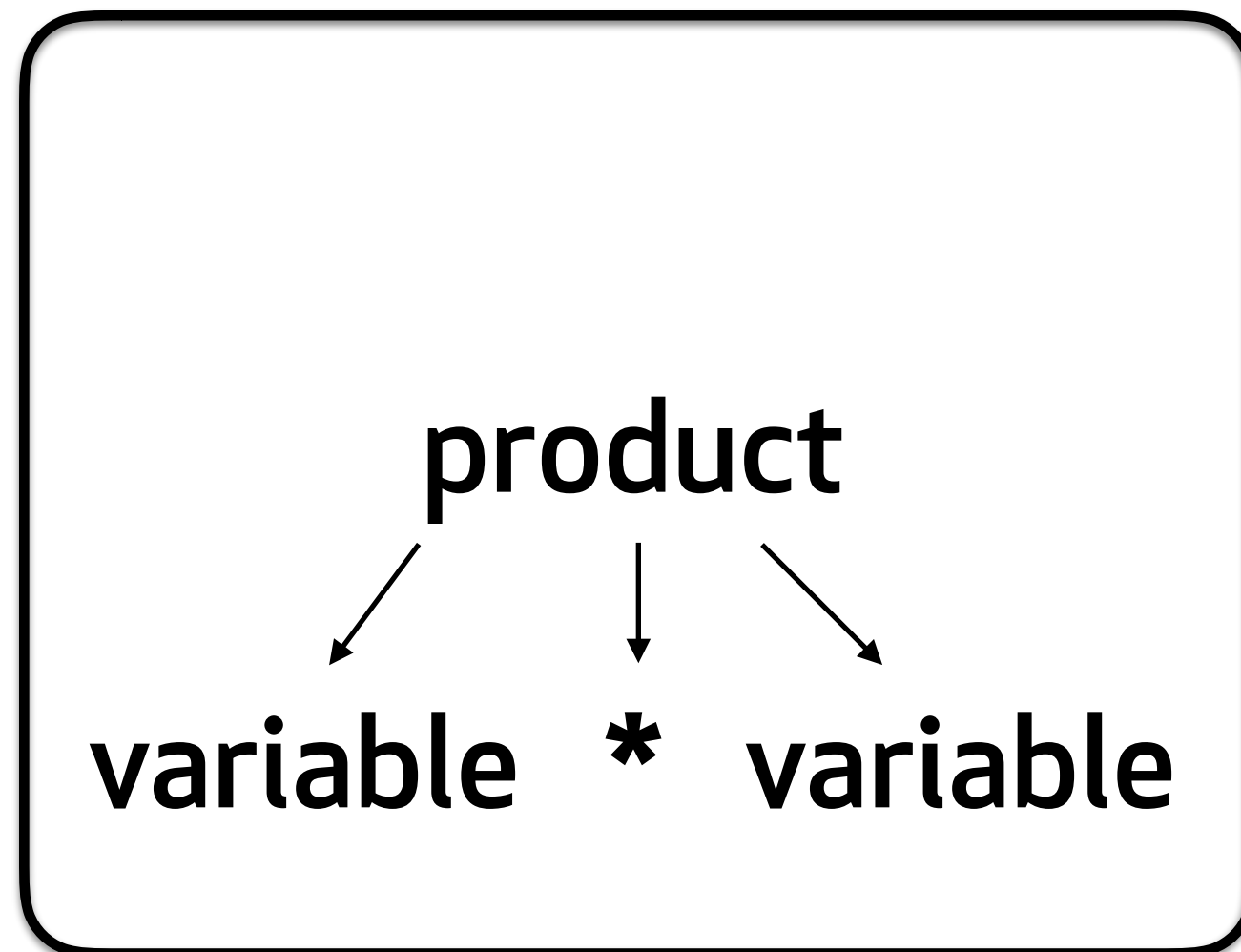


LR Parsing

Text: x * y + z



Stack:

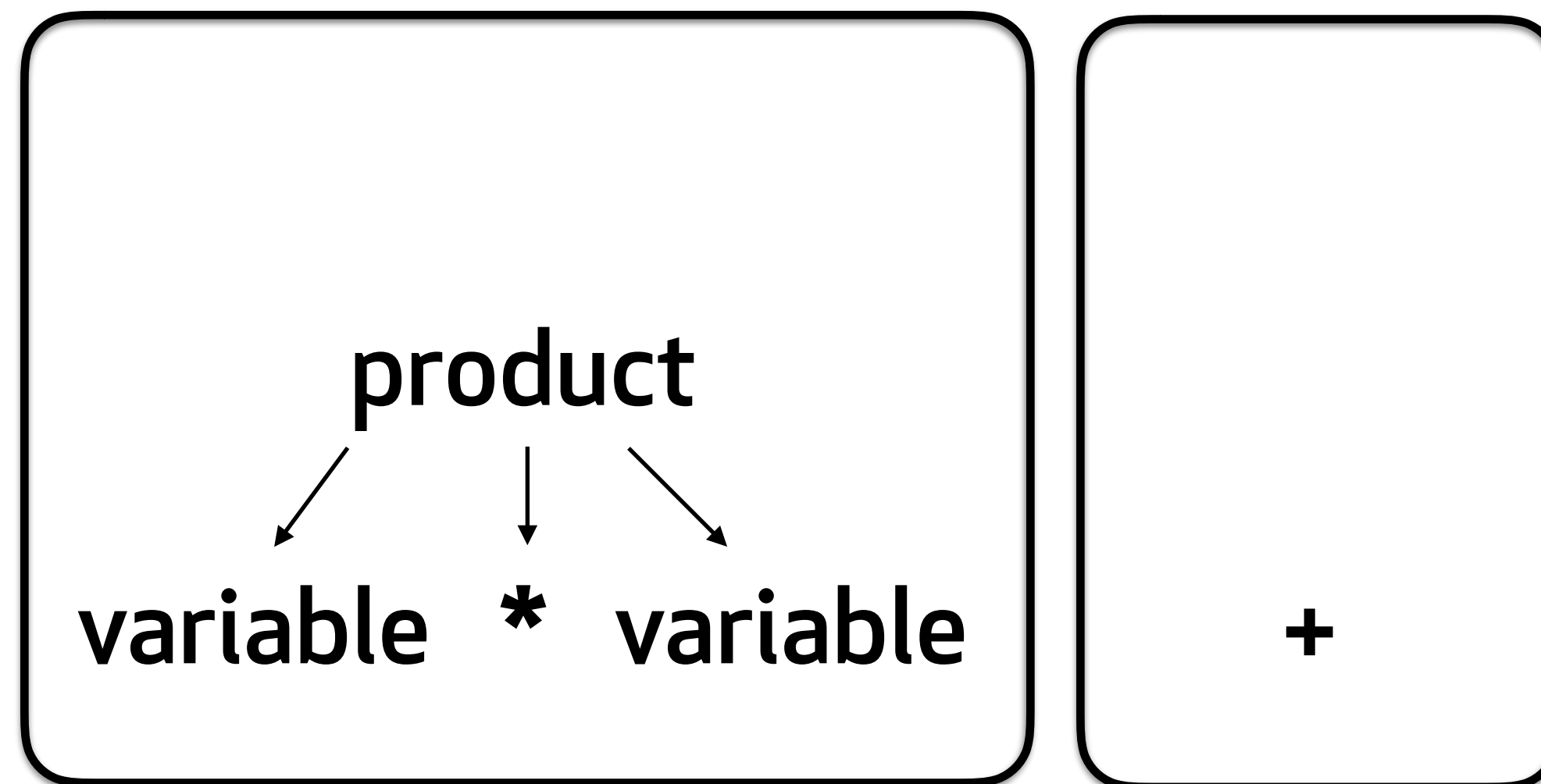


LR Parsing

Text: $x * y + z$



Stack:

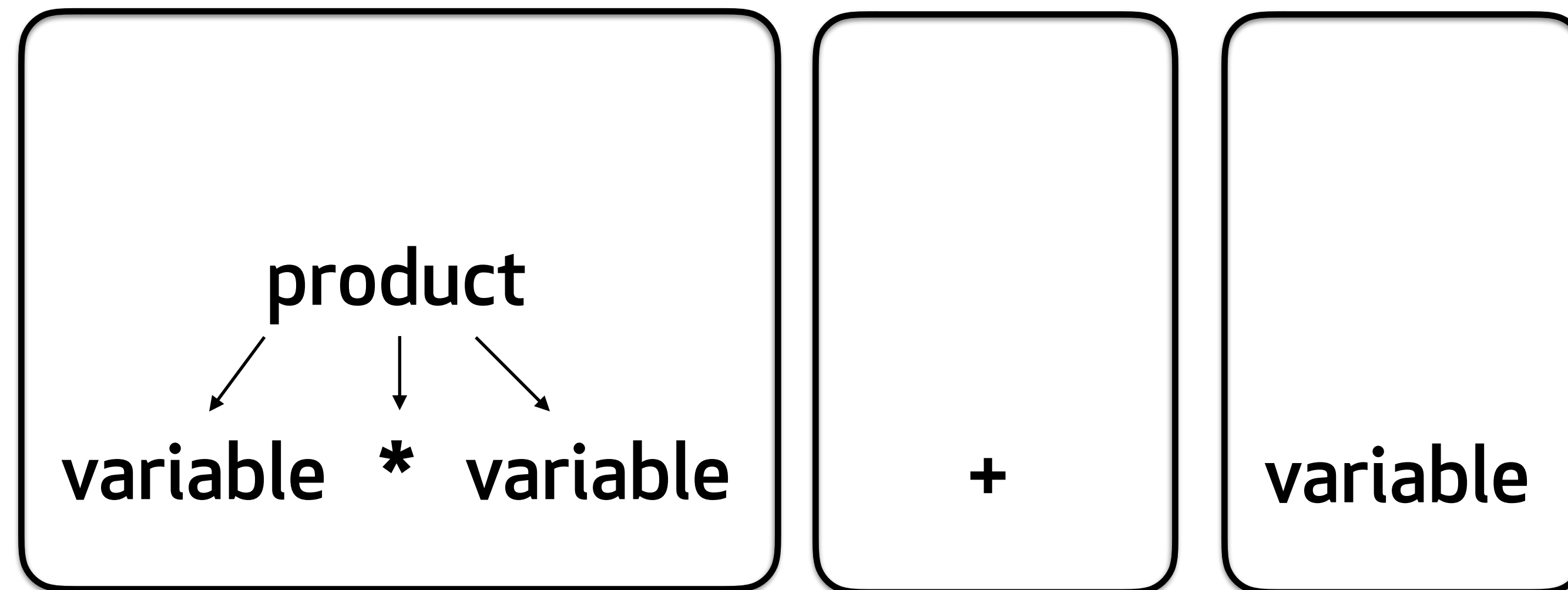


LR Parsing

Text: $x * y + z$



Stack:

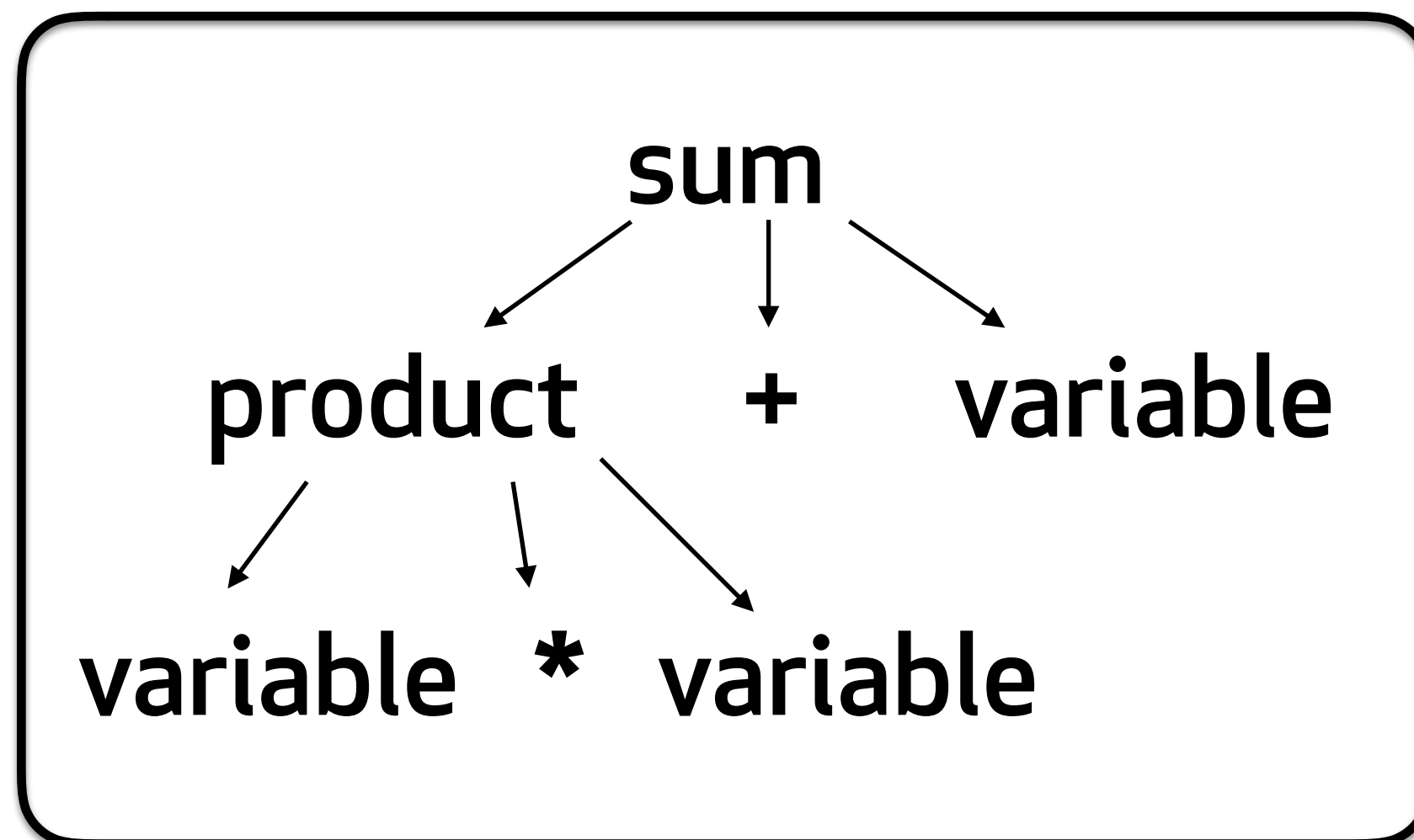


LR Parsing

Text: $x * y + z$



Stack:



LR Parsing

- Does this work for all languages?



LR Parsing

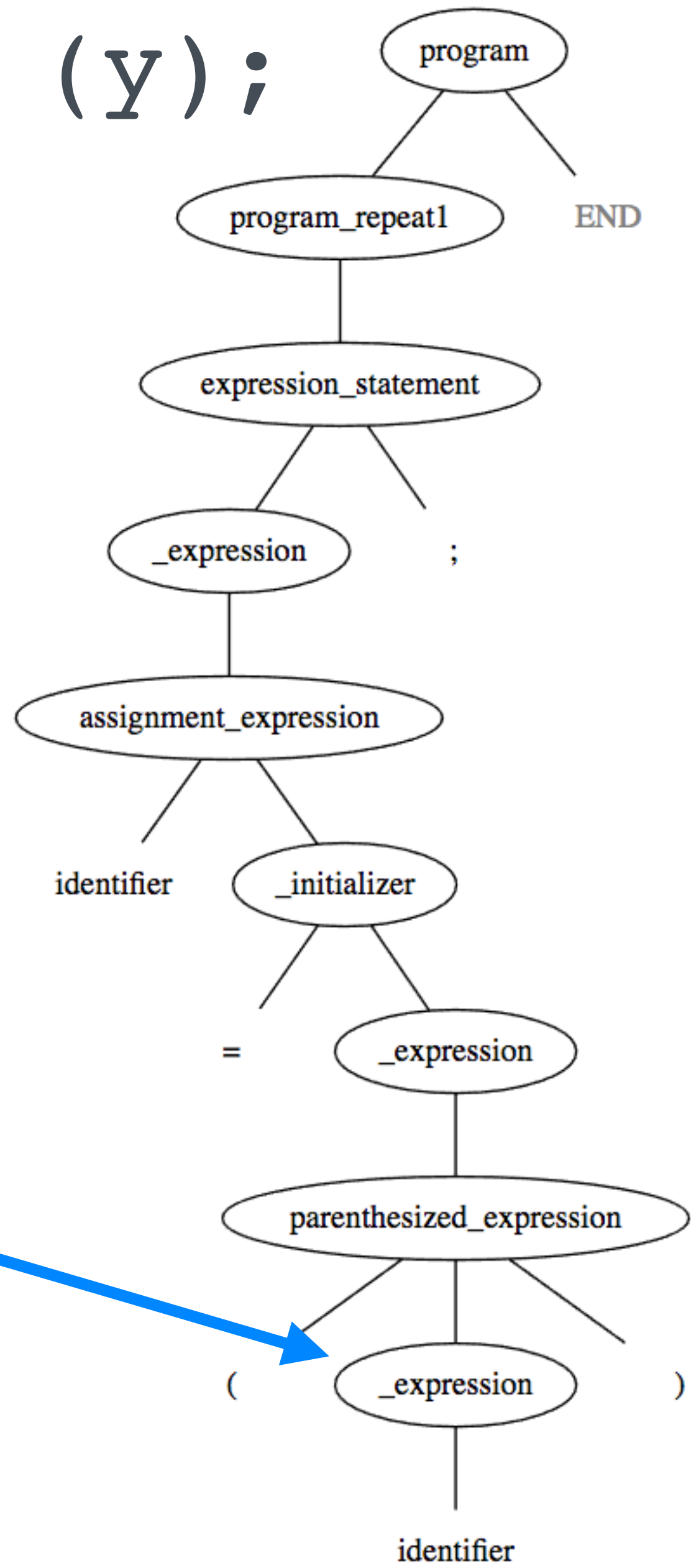
- Does this work for all languages? Nope.

```
x = (y);           // parenthesized expression
```

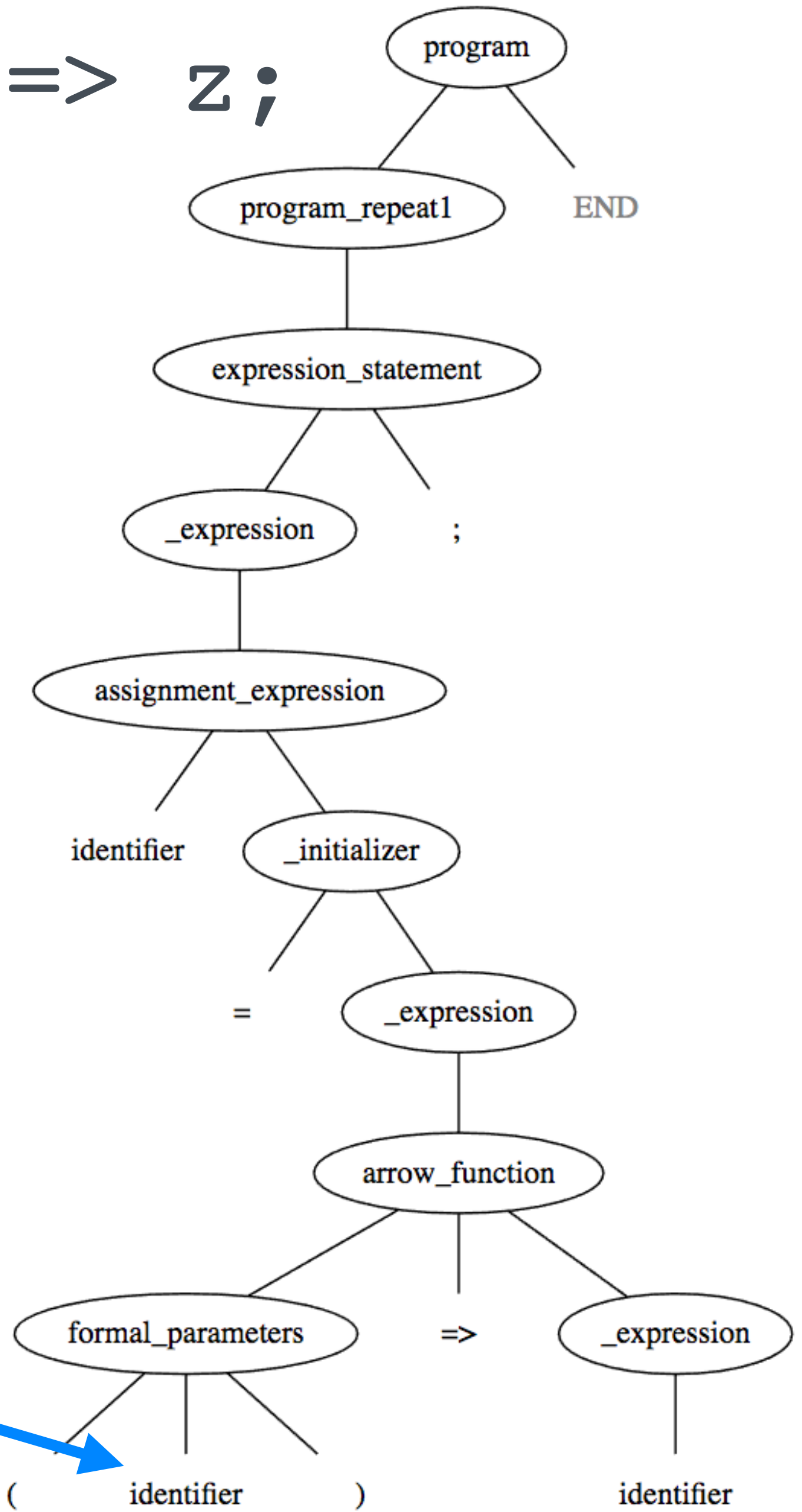
```
x = (y) => z;      // arrow function
```



$x = (y);$

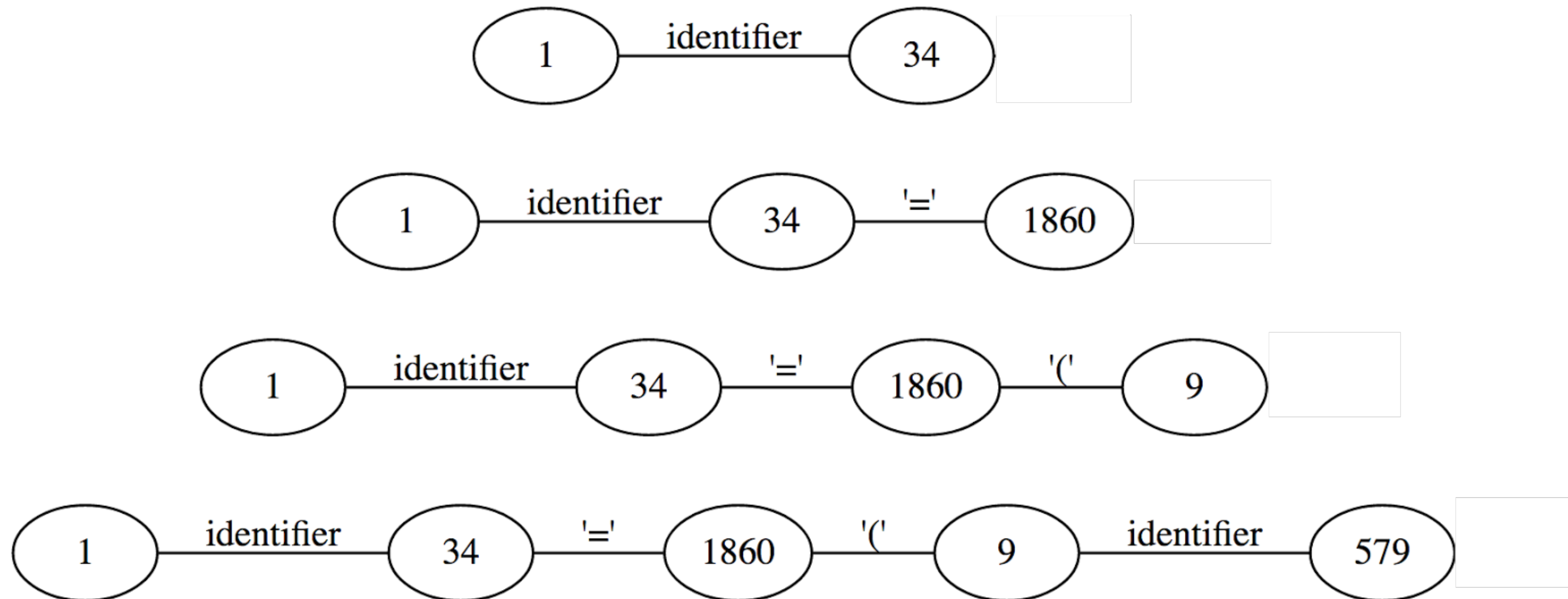


$x = (y) \Rightarrow z;$



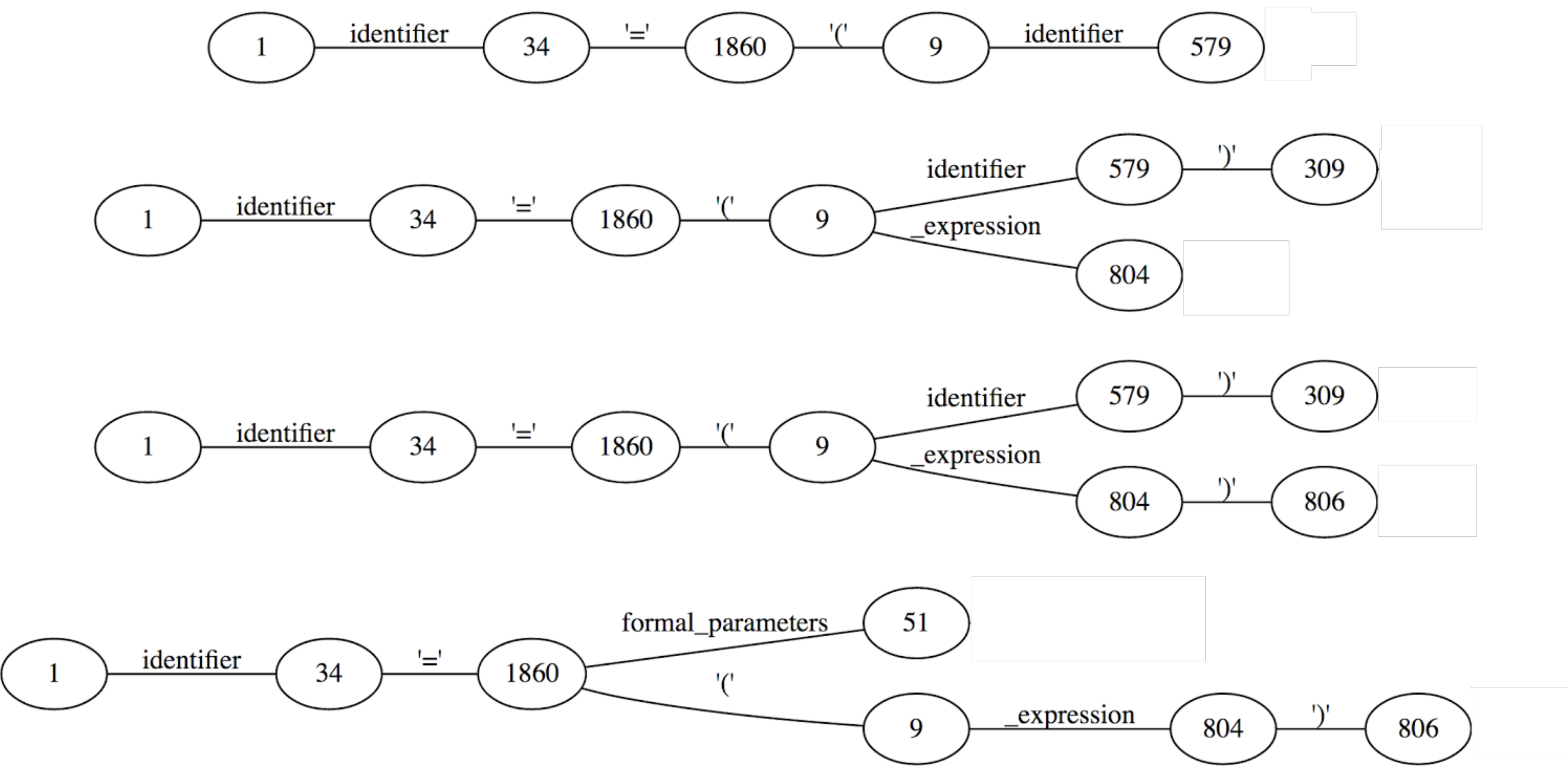
GLR Parsing

`x = (y) => z ;`



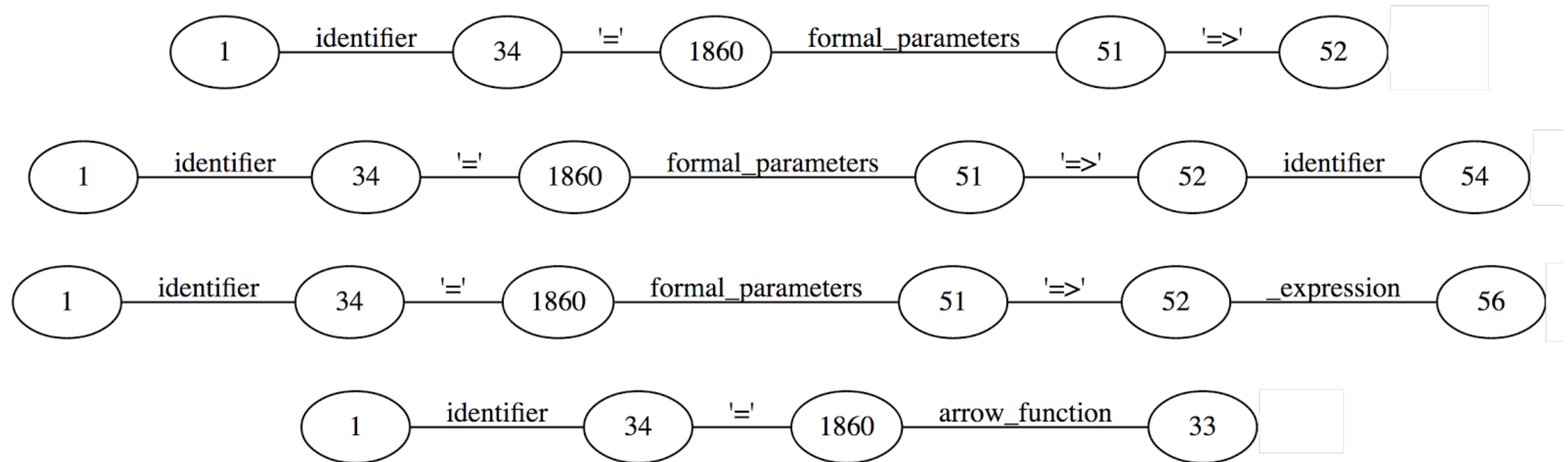
GLR Parsing

`x = (y) => z;`



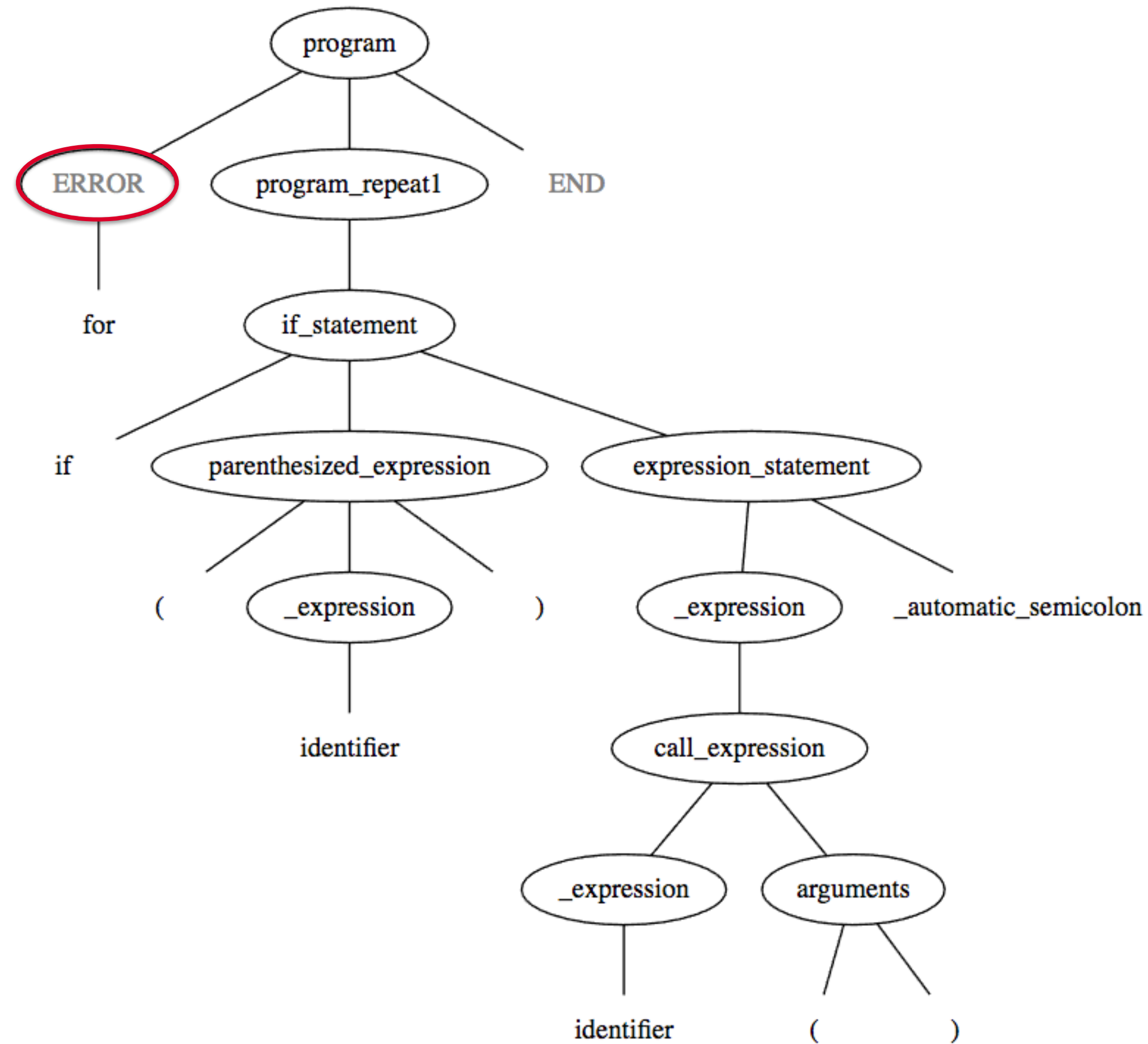
GLR Parsing

`x = (y) => z ;`



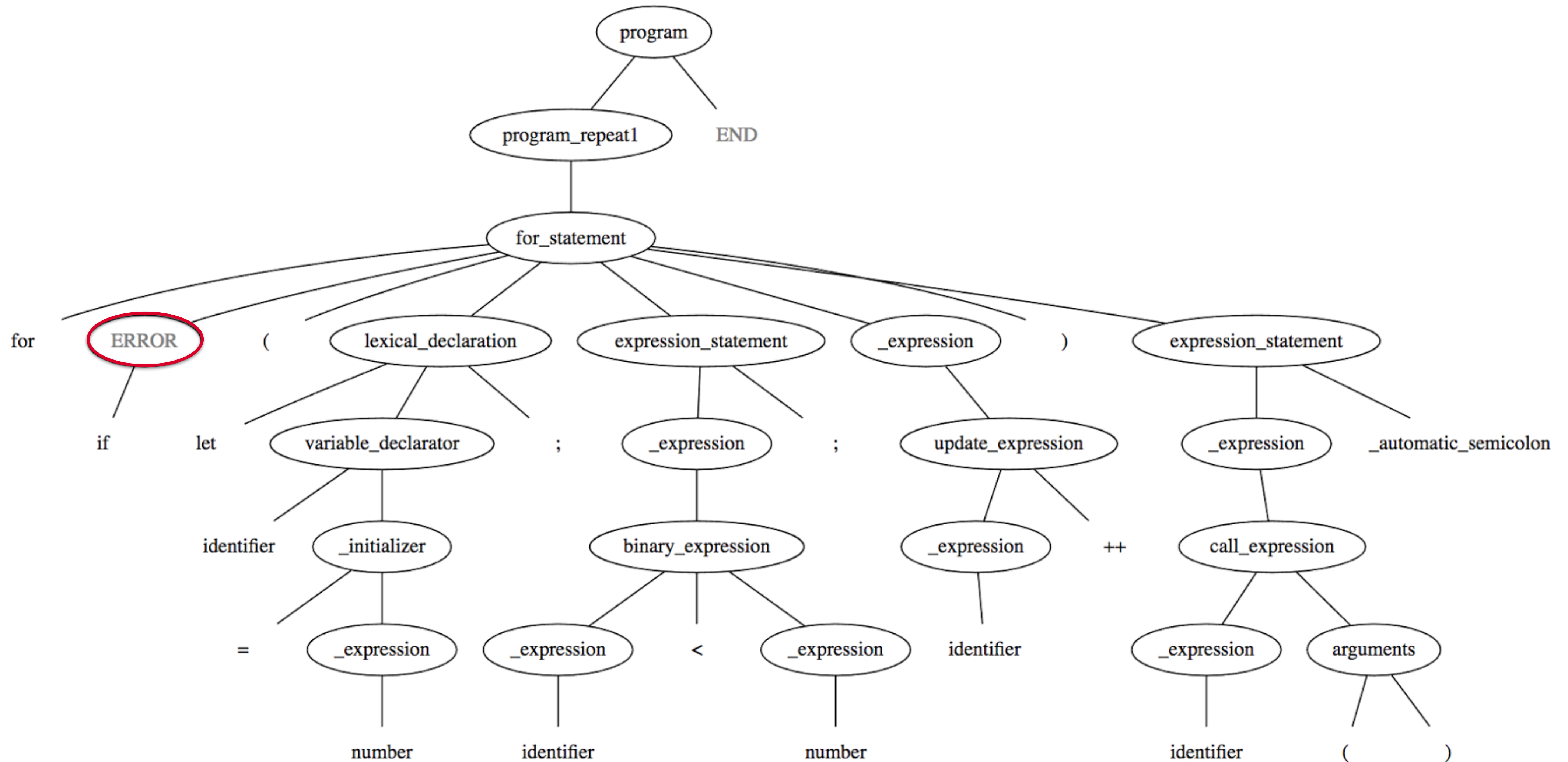
Error Recovery

for if (x) y()



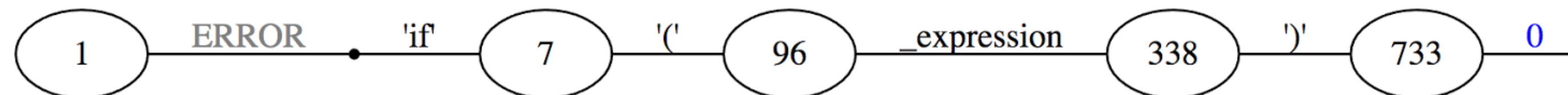
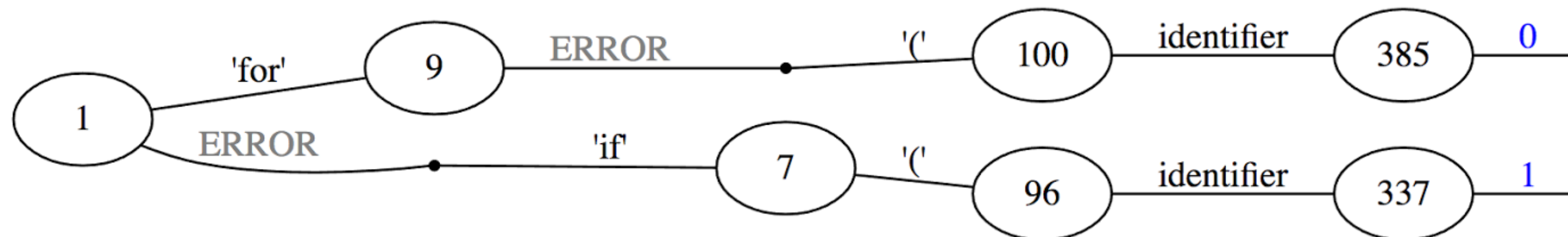
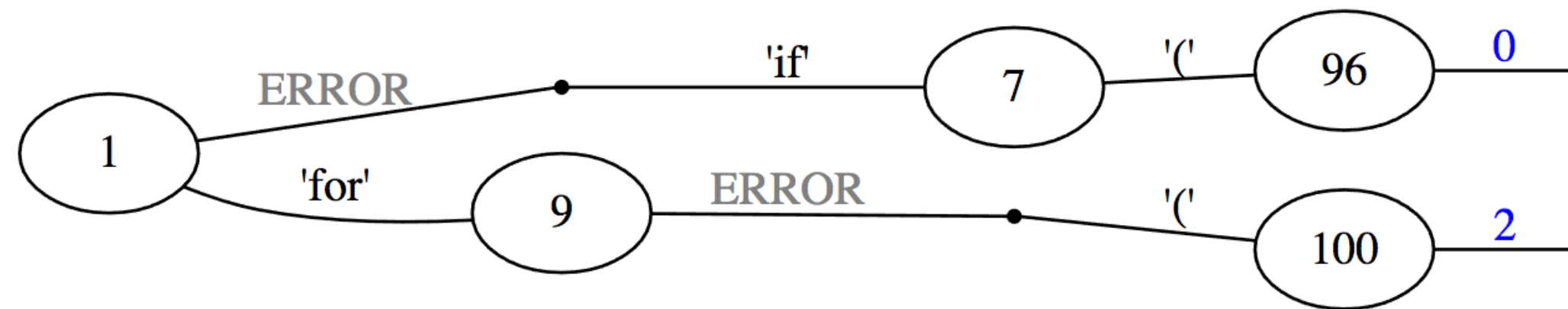
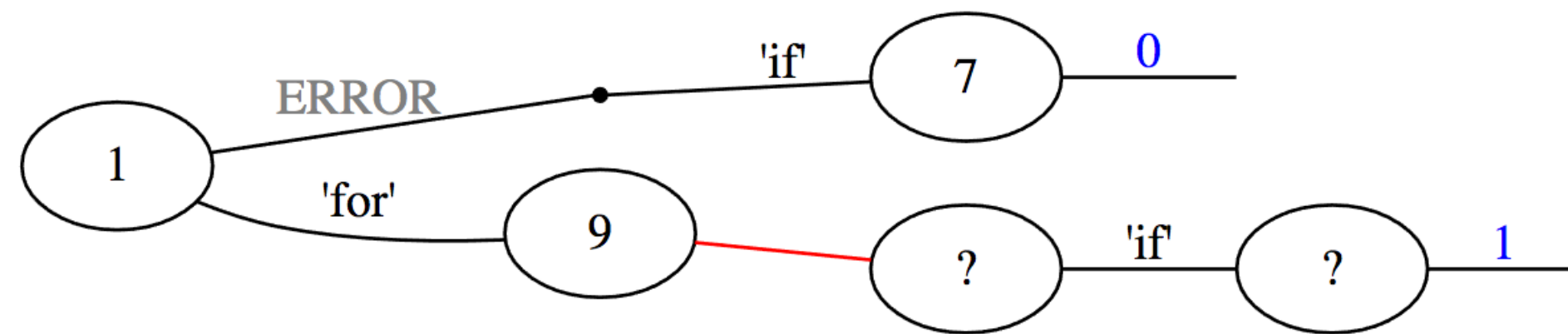
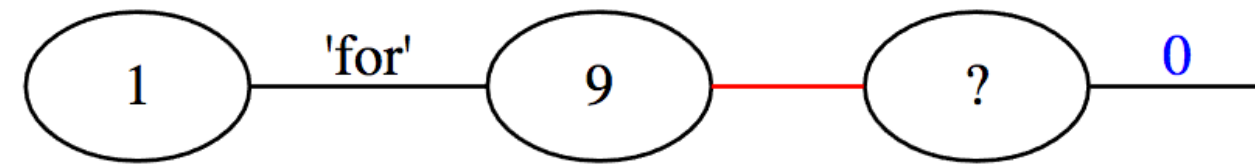
Error Recovery

for if (let x = 0; x < 5; x++) y()



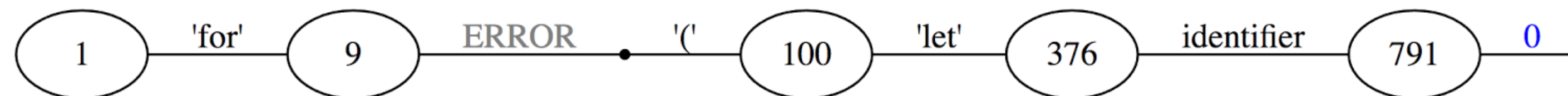
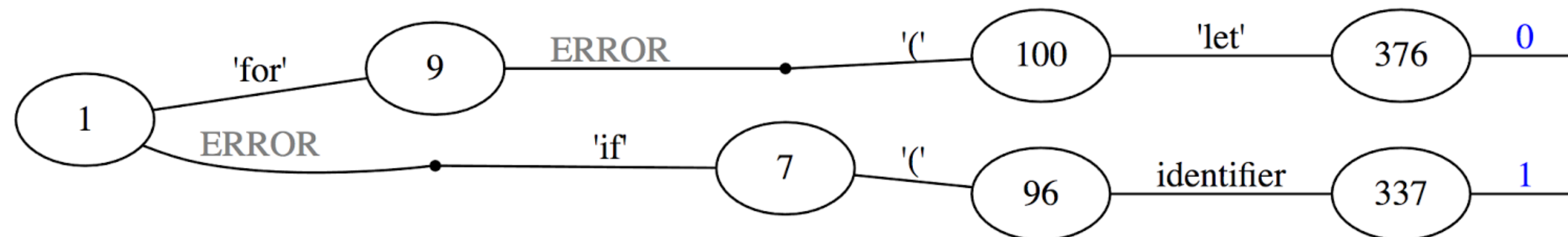
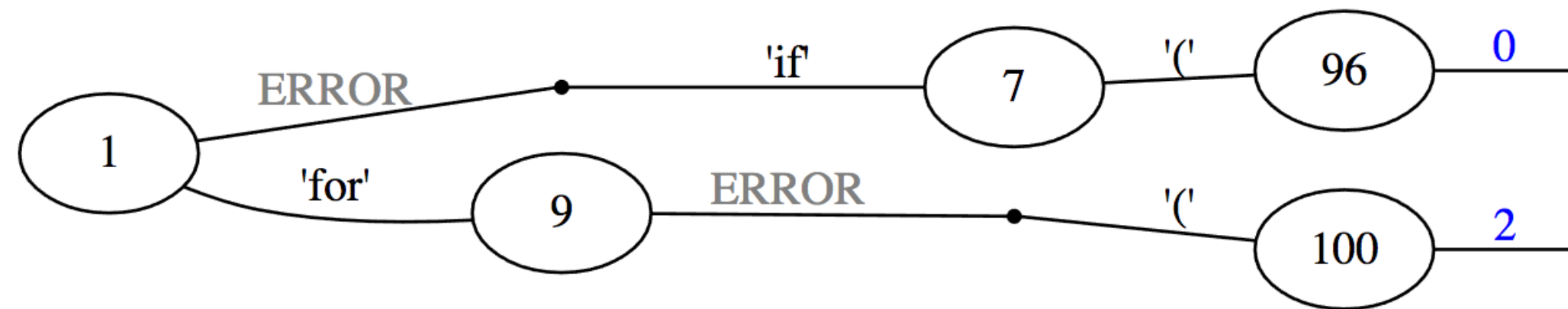
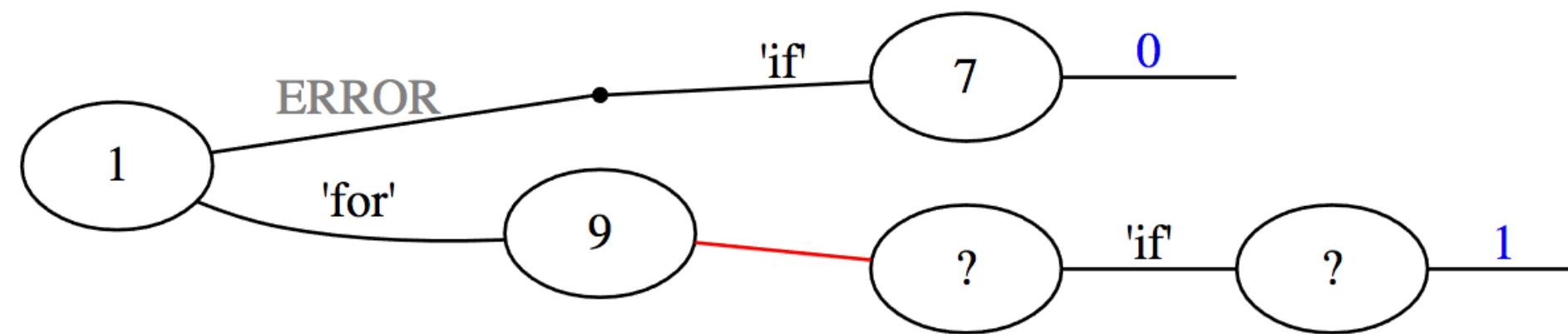
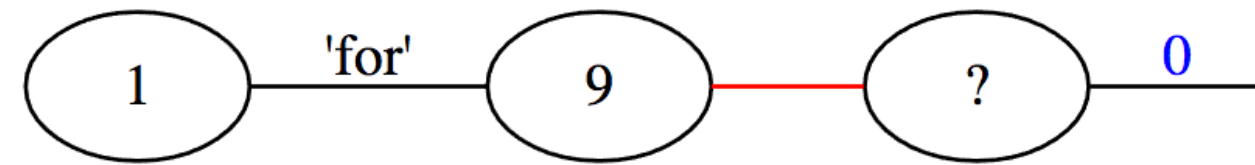
Error Recovery

for if (x) y()



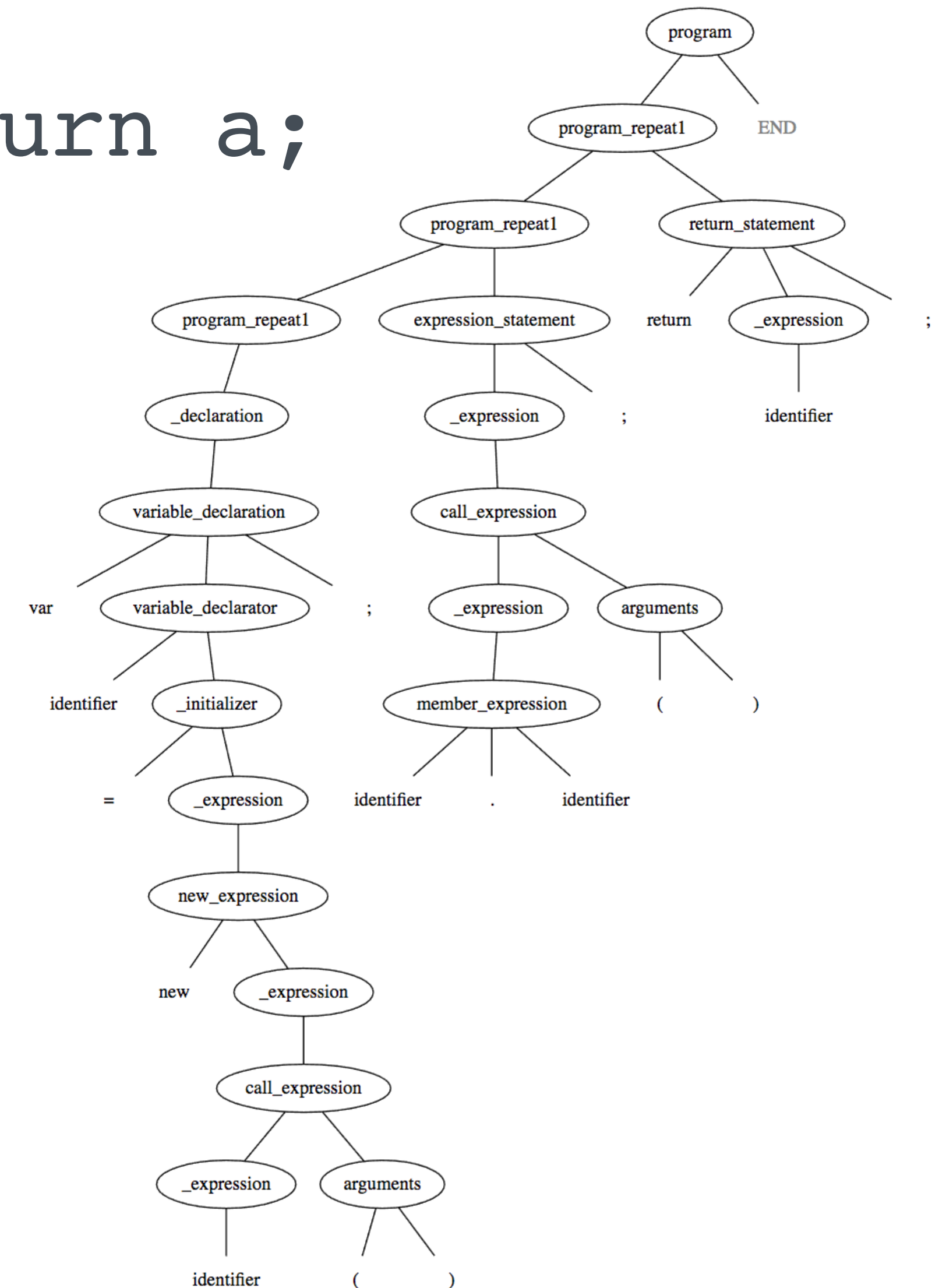
Error Recovery

```
for if (let x = 0; x < 5; x++) y()
```



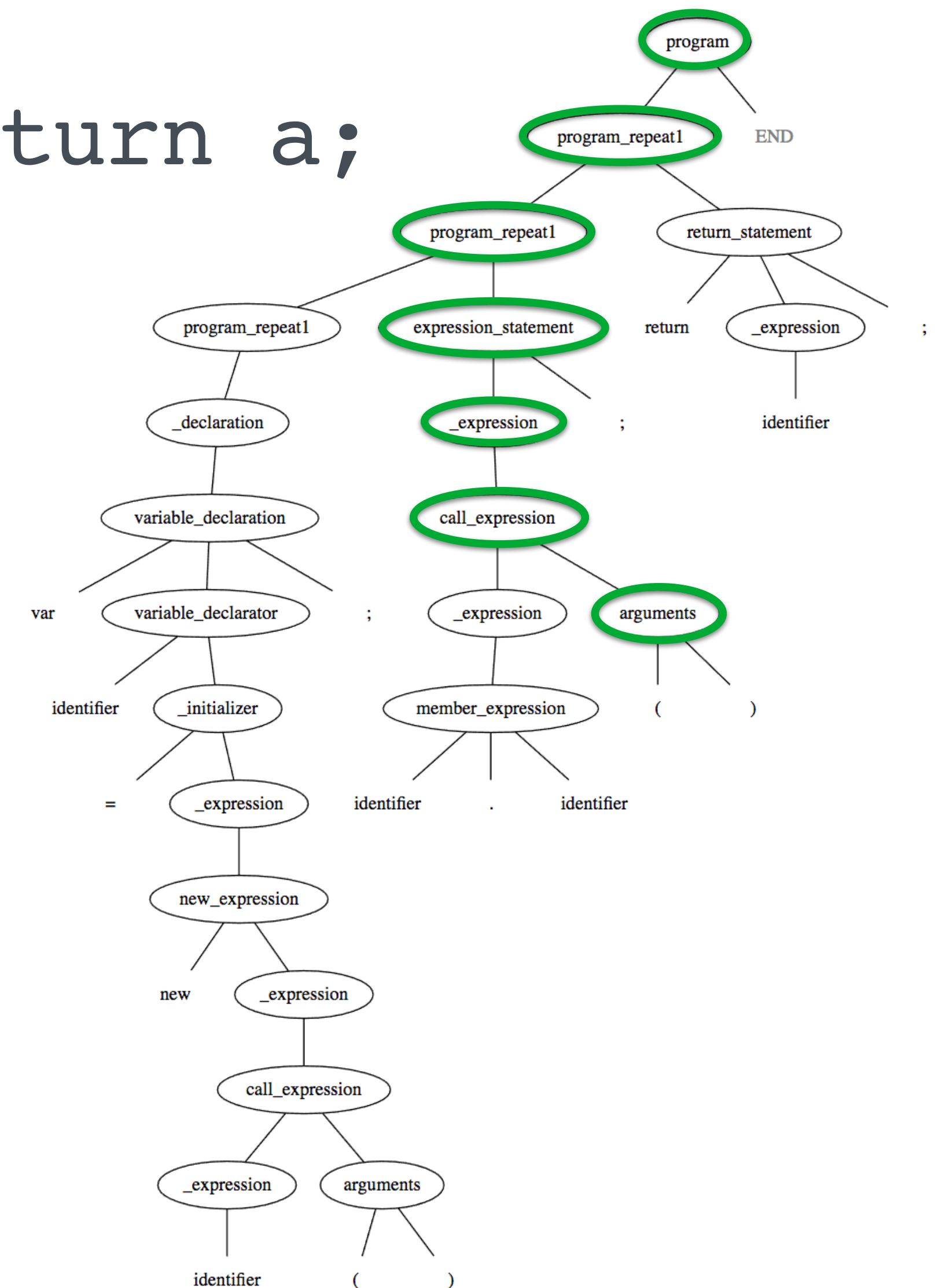
Incremental Parsing

Text: `var a = new B(); a.c(); return a;`



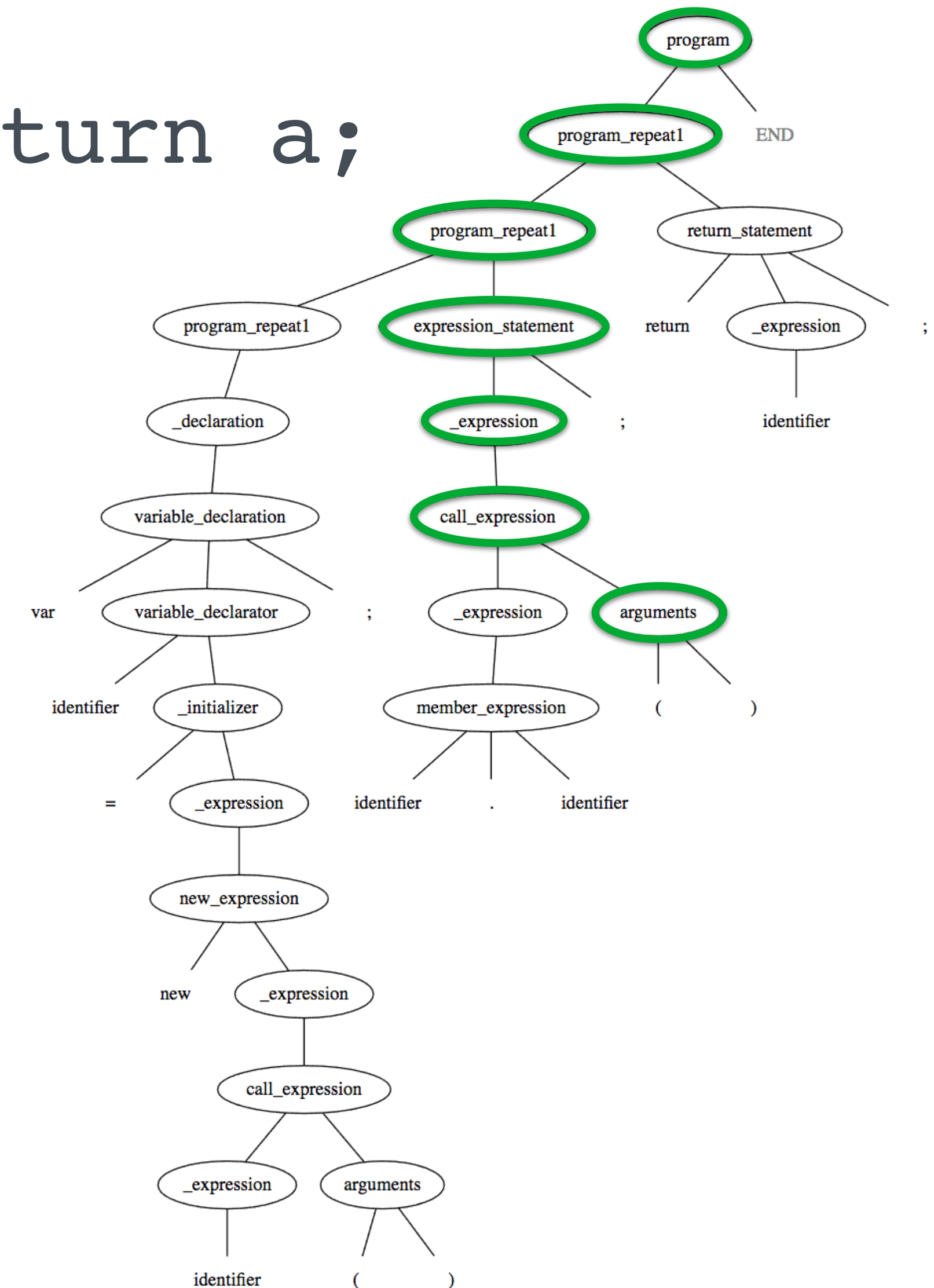
Incremental Parsing

Text: var a = new B(); a.c(d); return a;



Incremental Parsing

Text: `var a = new B(); a.c(d); return a;`



Stack:



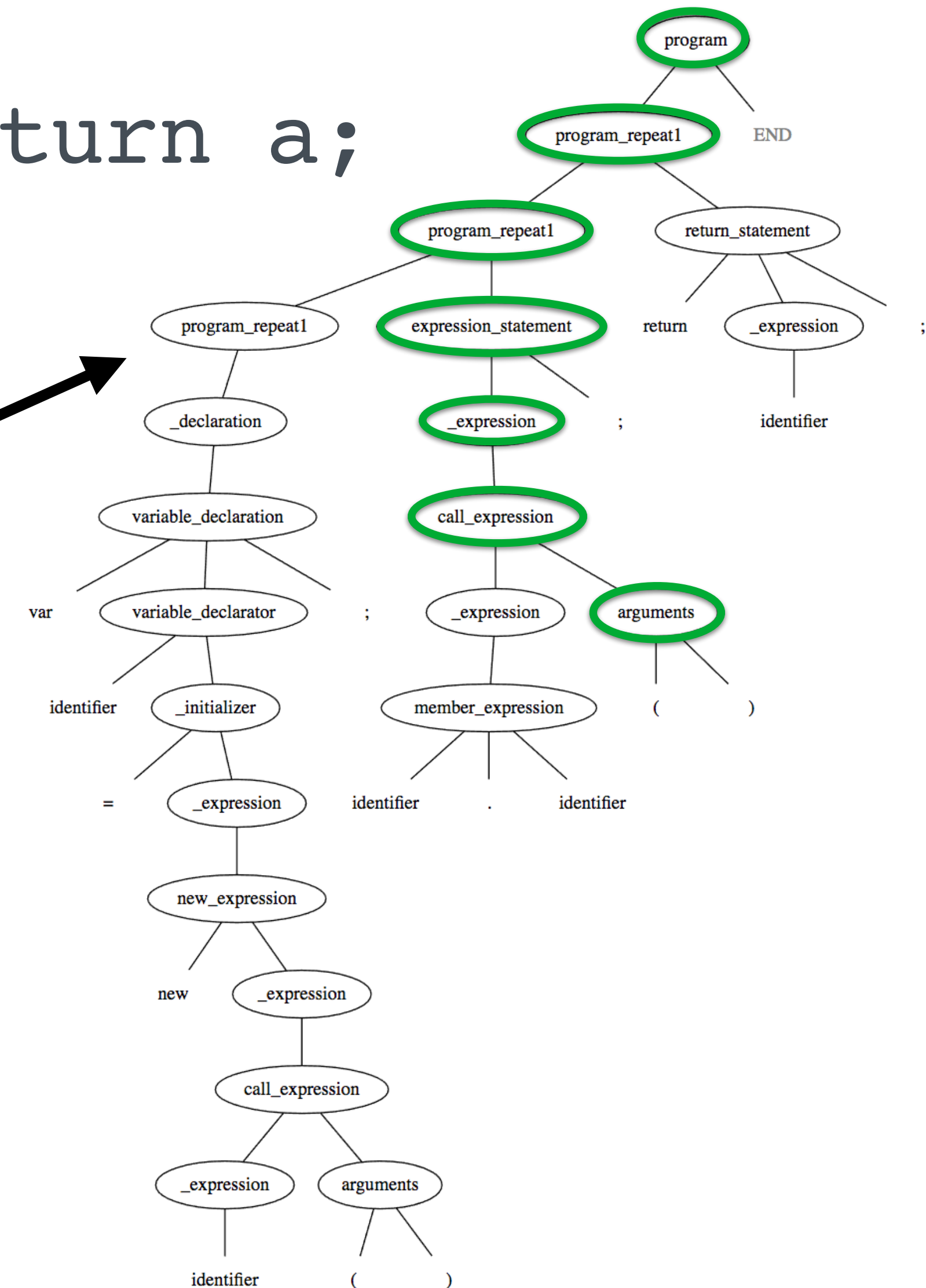
Incremental Parsing

Text: `var a = new B(); a.c(d); return a;`



Stack:

variable
declaration



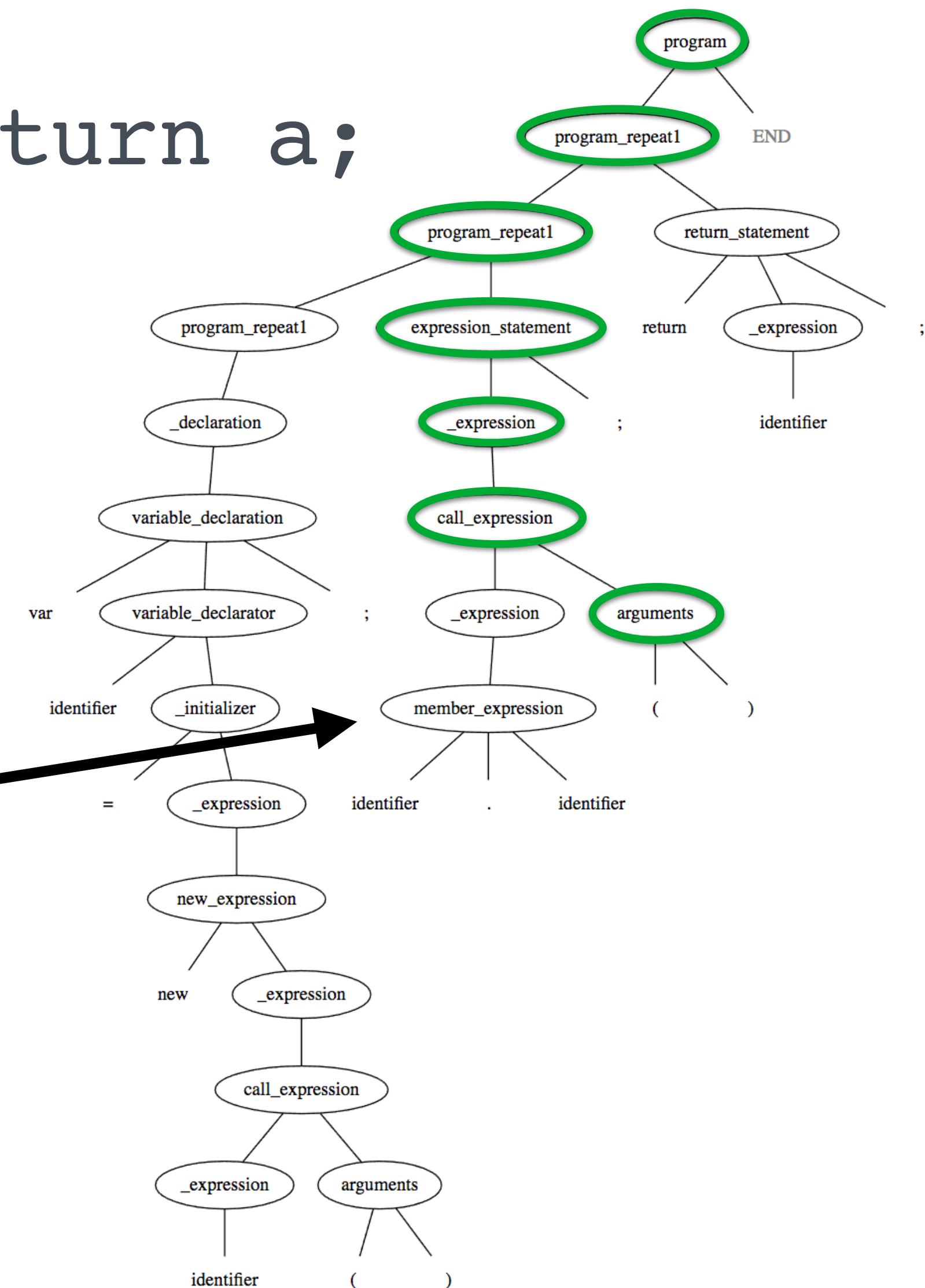
Incremental Parsing

Text: var a = new B(); a.c(**d**); return a;

Stack:

variable
declaration

member
expression



Incremental Parsing

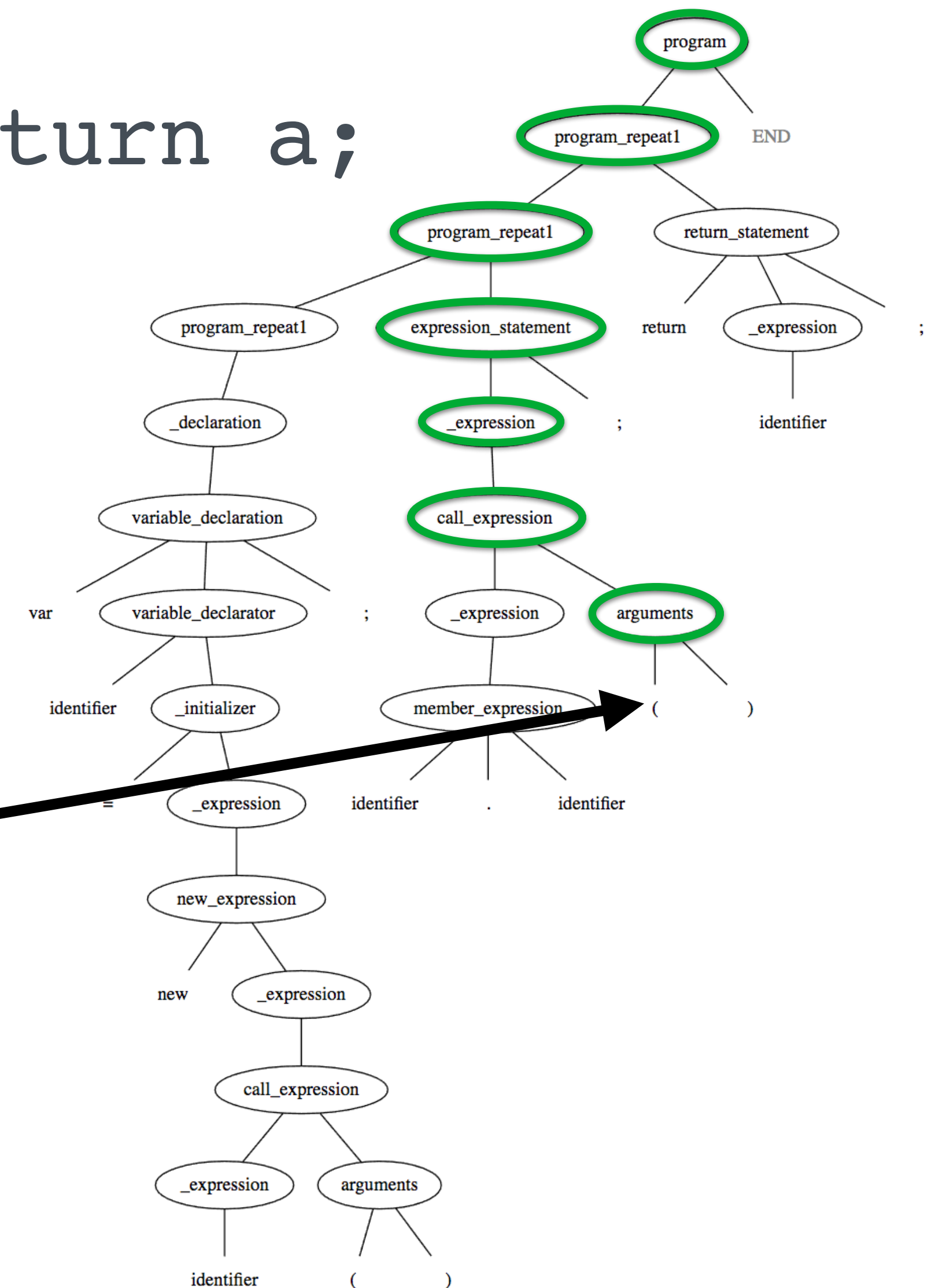
Text: `var a = new B(); a.c(d); return a;`

Stack:

variable
declaration

member
expression

'('



Incremental Parsing

Text: var a = new B(); a.c(**d**); return a;

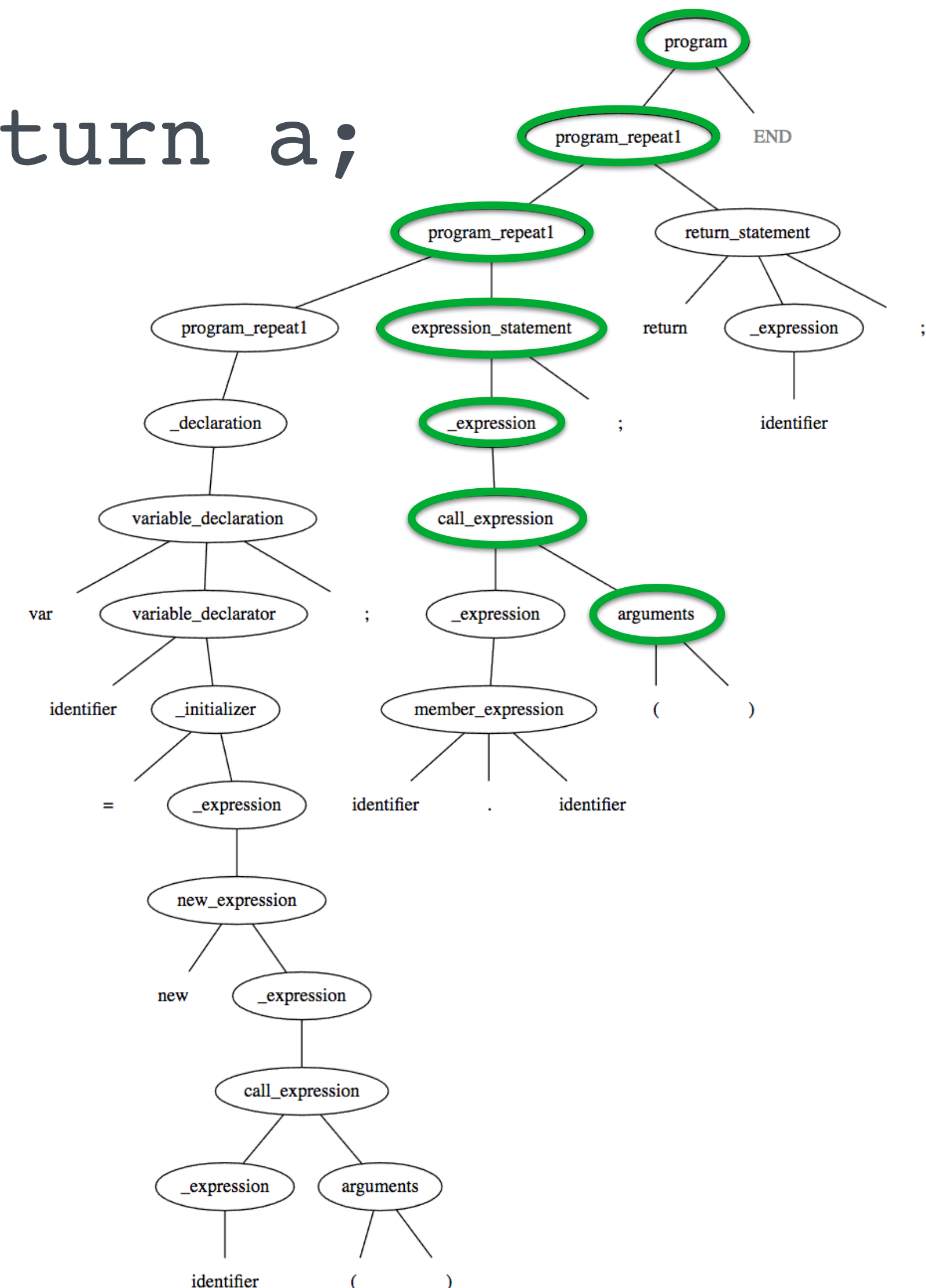
Stack:

variable
declaration

member
expression

'('

identifier



Incremental Parsing

Text: var a = new B(); a.c(**d**); return a;

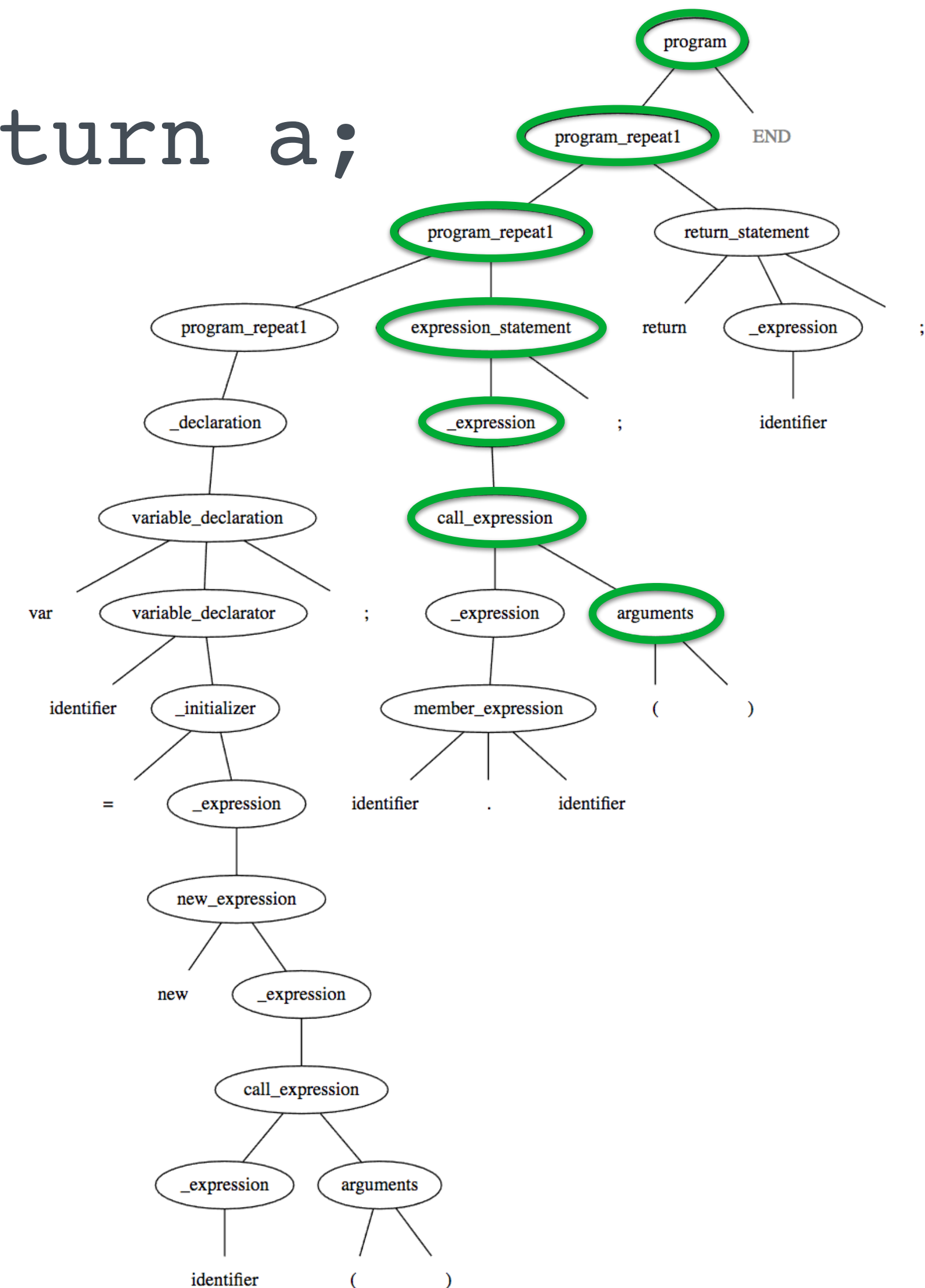
Stack:

variable
declaration

member
expression

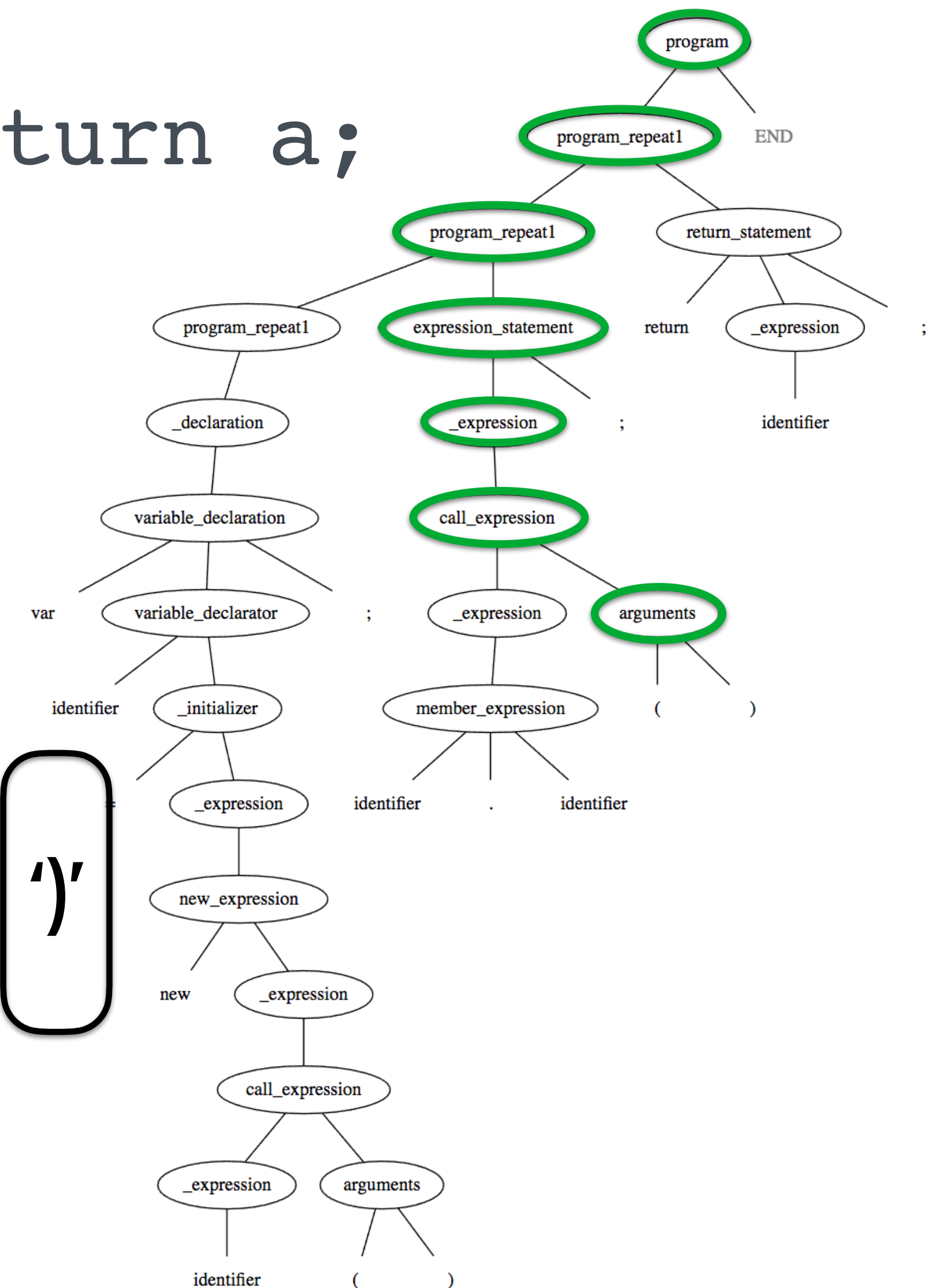
'('

expression



Incremental Parsing

Text: `var a = new B(); a.c(d); return a;`



Stack:

variable declaration

member expression

'(

expression

)

Incremental Parsing

Text: var a = new B(); a.c(**d**); return a;

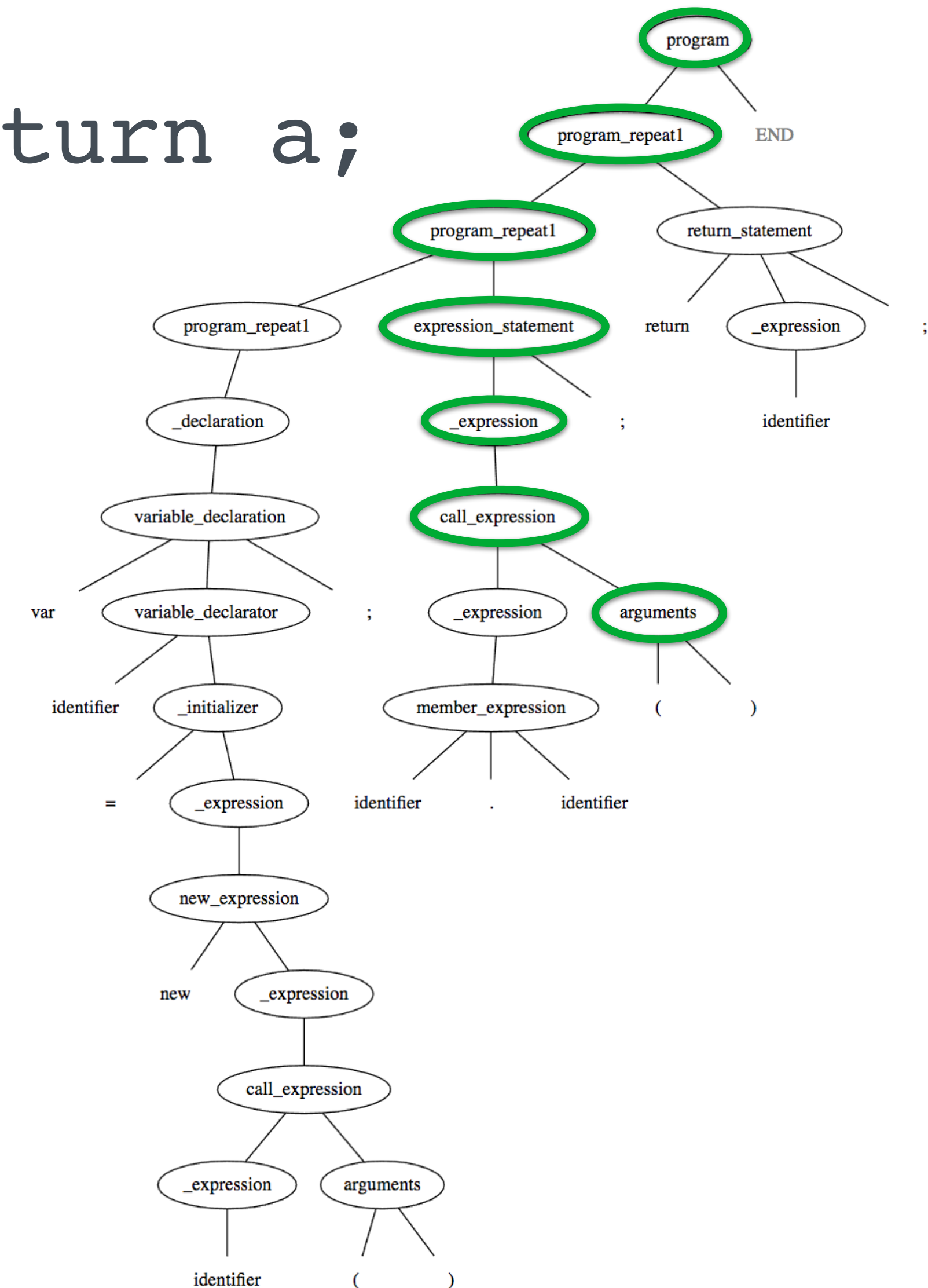


Stack:

variable
declaration

member
expression

arguments



Incremental Parsing

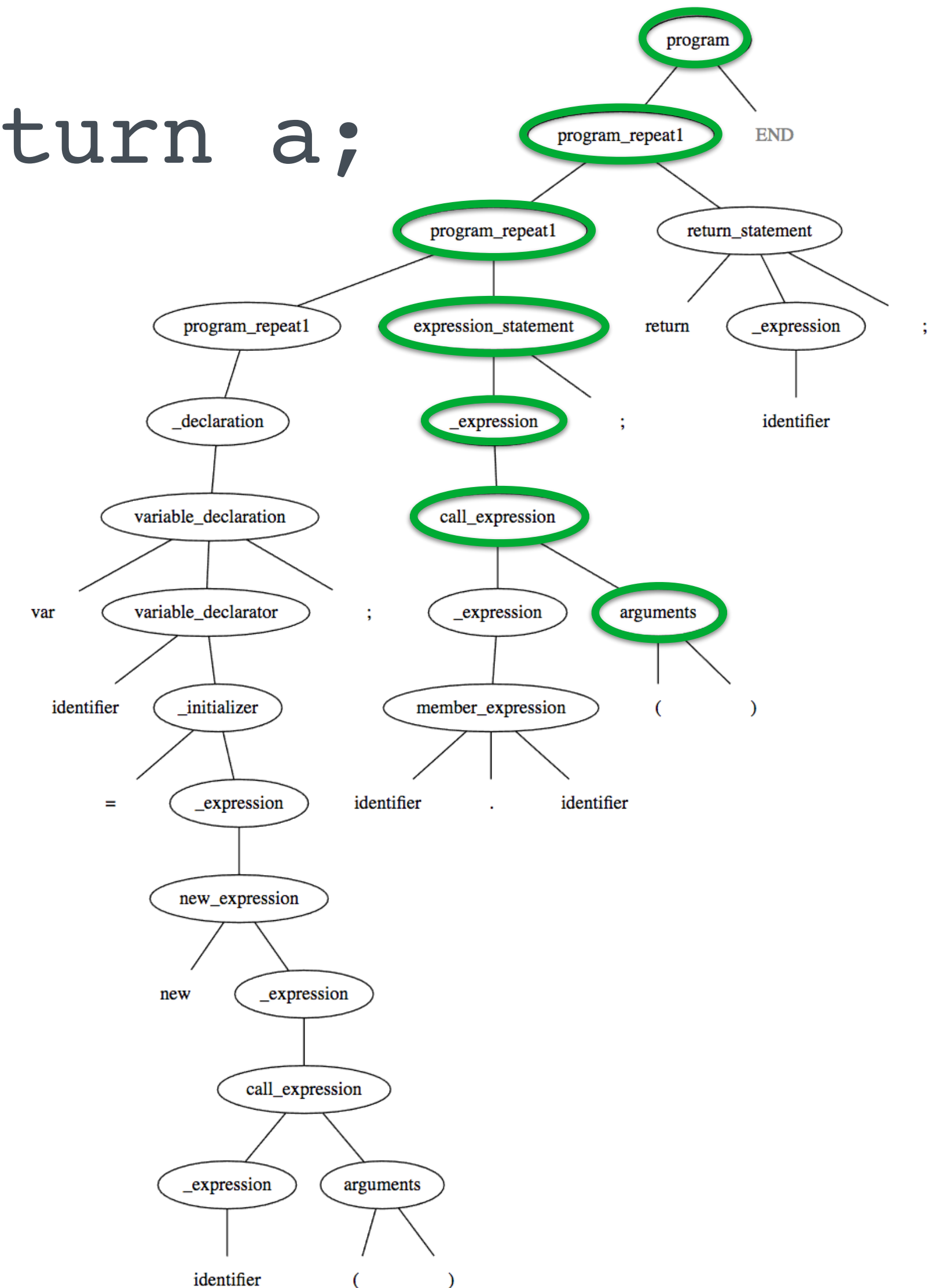
Text: var a = new B(); a.c(**d**); return a;



Stack:

variable
declaration

call
expression



Incremental Parsing

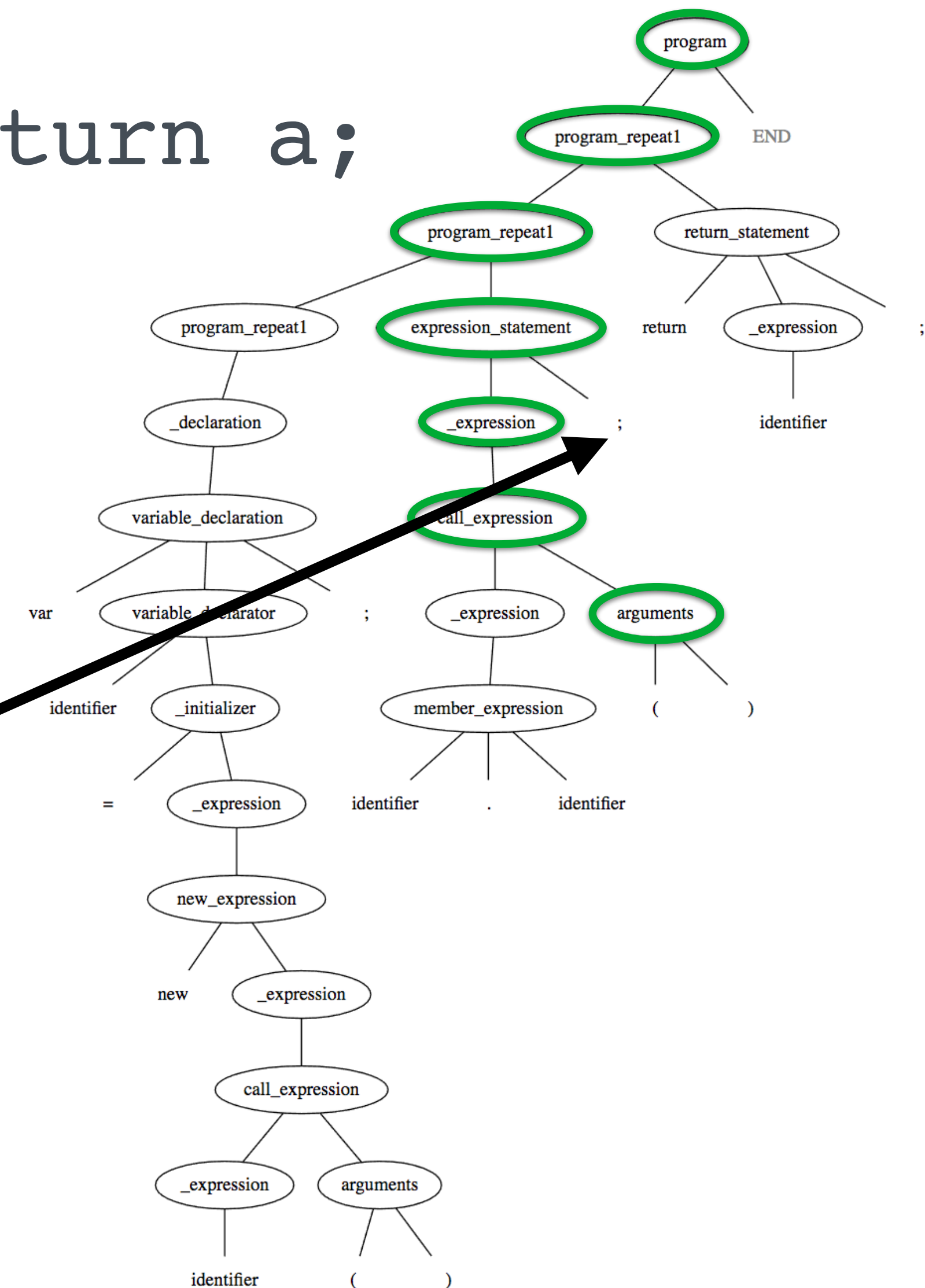
Text: `var a = new B(); a.c(d); return a;`

Stack:

variable
declaration

call
expression

;



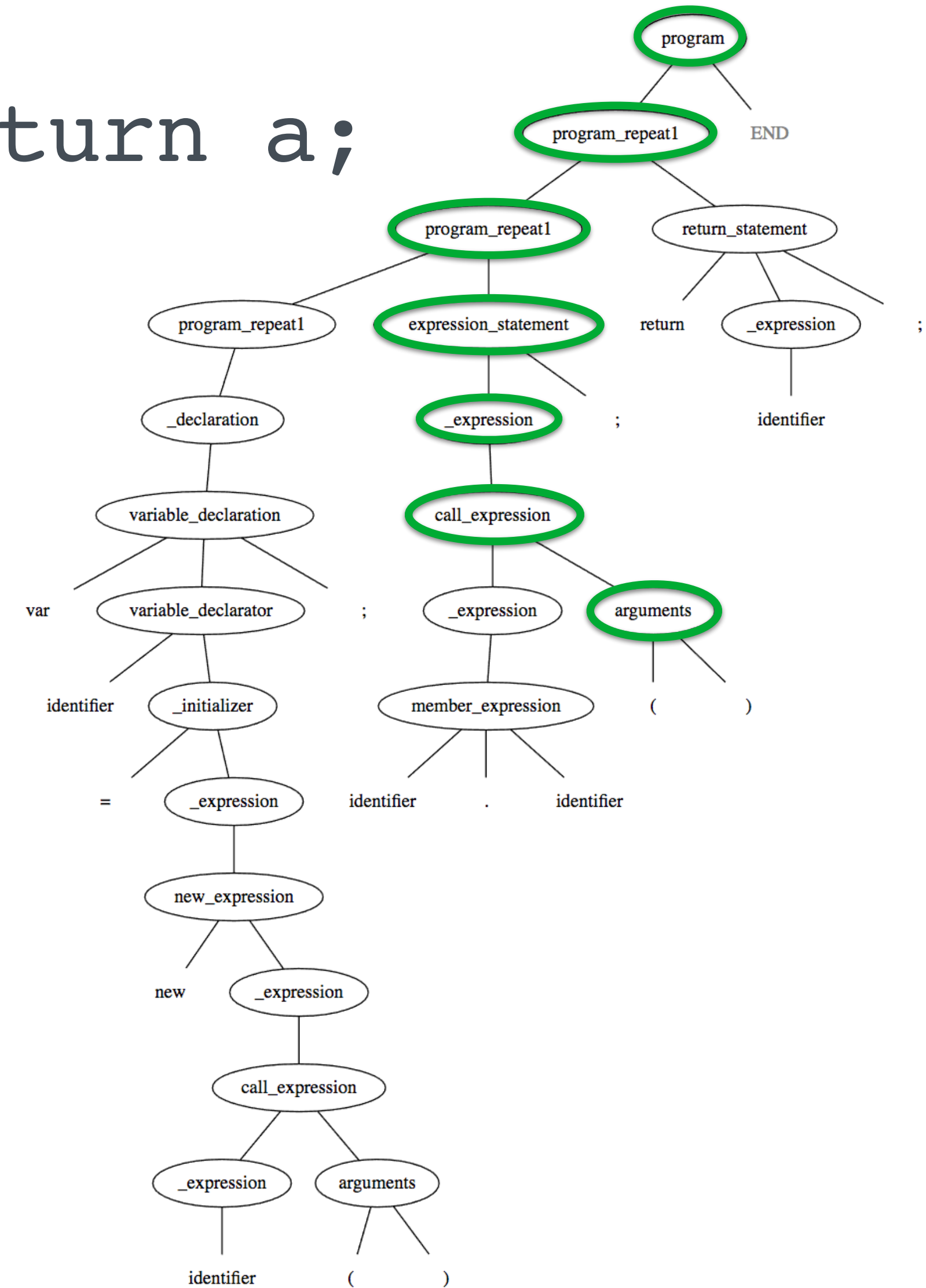
Incremental Parsing

Text: `var a = new B(); a.c(d); return a;`

Stack:

variable declaration

expression statement

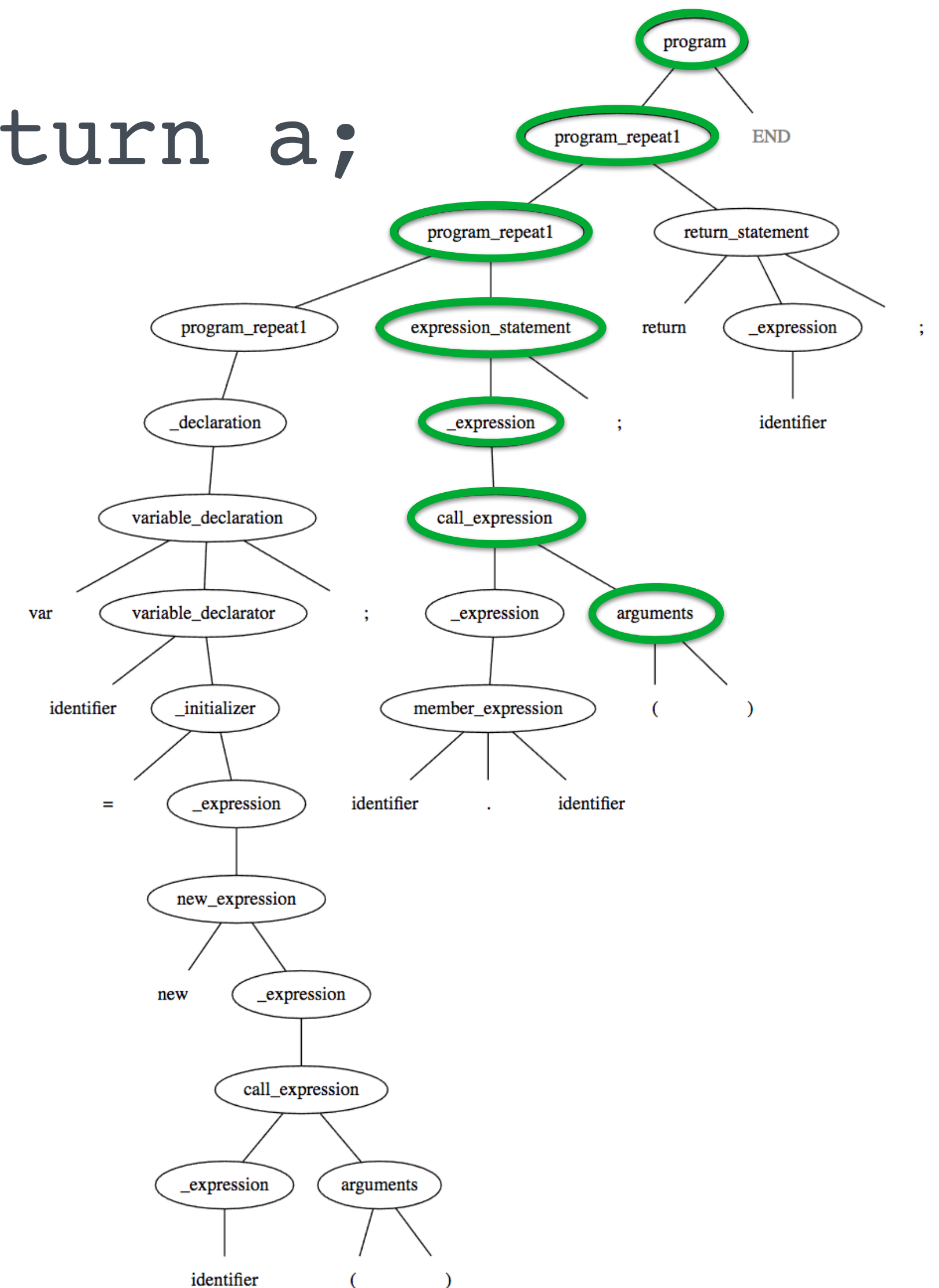


Incremental Parsing

Text: `var a = new B(); a.c(d); return a;`

Stack:

program repetition



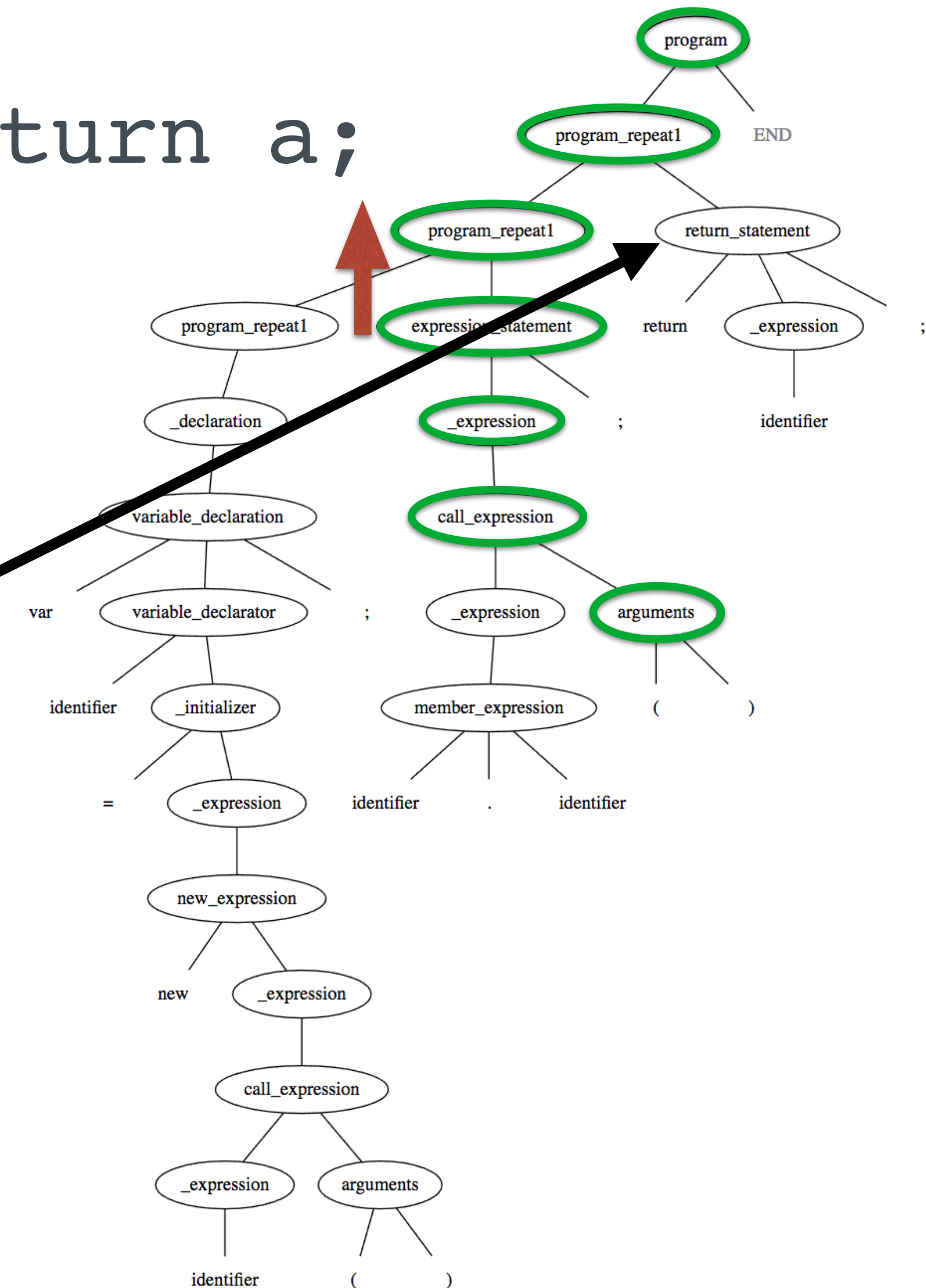
Incremental Parsing

Text: `var a = new B(); a.c(d); return a;`

Stack:

program
repetition

return
statement

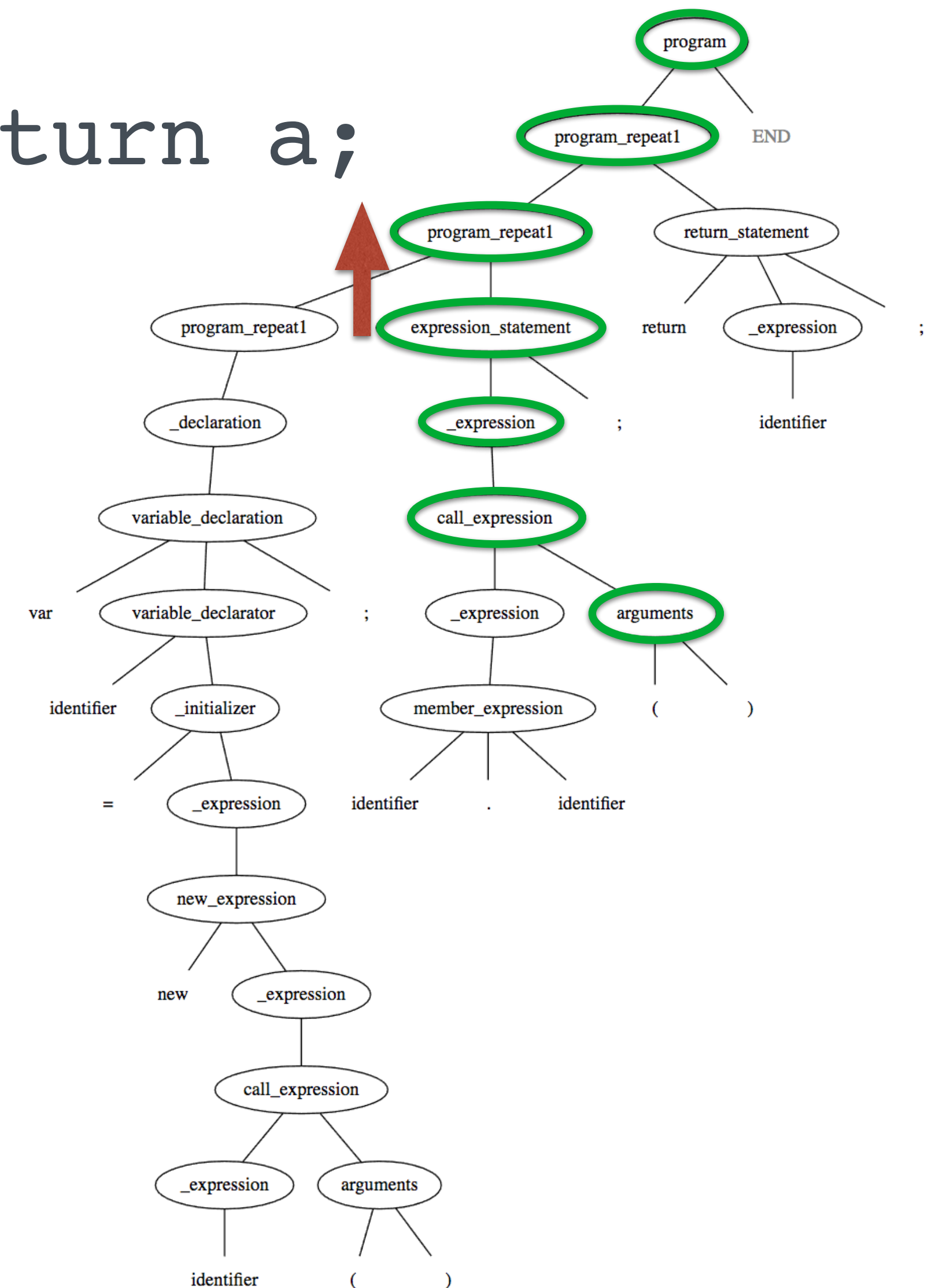


Incremental Parsing

Text: var a = new B(); a.c(**d**); return a;

Stack:

program



Next Steps



- Add support for more languages
- New Atom features using the syntax trees
- New GitHub.com features using syntax trees



Thank You!

@maxbrunsfeld

