Follow sets. LL(1) Parsing Table

Exercise Introducing Follow Sets

Compute nullable, first for this grammar:

stmtList ::= ε | stmt stmtList

stmt ::= assign | block

assign ::= **ID = ID ;**

block ::= **beginof ID** stmtList **ID ends**

Describe a parser for this grammar and explain how it behaves on this input:

beginof myPrettyCode

x = u;

y = v;

myPrettyCode **ends**

How does a recursive descent parser look like?

```
def stmtList =
  if (???) {}                what should the condition be?
  else { stmt; stmtList }
def stmt =
  if (lex.token == ID) assign
  else if (lex.token == beginof) block
  else error("Syntax error: expected ID or beginonf")
...
def block =
  { skip(beginof); skip(ID); stmtList; skip(ID); skip(ends) }
```

Problem Identified

stmtList ::= ε | stmt stmtList

stmt ::= assign | block

assign ::= **ID = ID ;**

block ::= **beginof ID** stmtList **ID ends**

Problem parsing stmtList:

- **ID** could start alternative stmt stmtList
- **ID** could **follow** stmt, so we may wish to parse **ε** that is, do nothing and return

- For nullable non-terminals, we must also compute what **follows** them

LL(1) Grammar - good for building recursive descent parsers

- Grammar is LL(1) if for each nonterminal X
  - first sets of different alternatives of X are disjoint
  - if nullable(X), first(X) must be disjoint from follow(X) and only one alternative of X may be nullable
- For each LL(1) grammar we can build recursive-descent parser
- Each LL(1) grammar is unambiguous
- If a grammar is not LL(1), we can sometimes transform it into equivalent LL(1) grammar

Computing if a token can **follow**

first$(B_1 \ldots B_p) = \{a \in \Sigma \mid B_1 \ldots B_p \Rightarrow \ldots \Rightarrow aw\}$

follow$(X) = \{a \in \Sigma \mid S \Rightarrow \ldots \Rightarrow \ldots Xa \ldots\}$

There exists a derivation from the start symbol that produces a sequence of terminals and nonterminals of the form ...Xa...
(the token a follows the non-terminal X)

Rule for Computing Follow

Given $X ::= YZ$ (for reachable X)

then **first**$(Z) \subseteq$ **follow**(Y)

and **follow**$(X) \subseteq$ **follow**(Z)

now take care of nullable ones as well:

For each rule $X ::= Y_1 \ldots Y_p \ldots Y_q \ldots Y_r$

follow(Y_p) should contain:

- **first**$(Y_{p+1} Y_{p+2} \ldots Y_r)$
- also **follow**(X) if **nullable**$(Y_{p+1} Y_{p+2} Y_r)$

Compute nullable, first, follow

stmtList ::= ε | stmt stmtList

stmt ::= assign | block

assign ::= **ID = ID ;**

block ::= **beginof ID** stmtList **ID ends**

Is this grammar LL(1)?

Conclusion of the Solution

The grammar is not LL(1) because we have

- nullable(stmtList)
- first(stmt) ∩ follow(stmtList) = {**ID**}

- If a recursive-descent parser sees **ID**, it does not know if it should
  - finish parsing stmtList or
  - parse another stmt

Table for LL(1) Parser: Example

S ::= B **EOF**

(1)

B ::= ε | B **(**B**)**

(1) **(2)**

nullable: B

first(S) = { **(**, **EOF** }

follow(S) = {}

first(B) = { **(** }

follow(B) = { **)**, **(**, **EOF** }

empty entry:
when parsing S,
if we see) ,
report error

Parsing table:

	EOF	(	)
S	{1}	{1}	{}
B	{1}	{1,2}	{1}

parse conflict - choice ambiguity: grammar not LL(1)

1 is in entry because **(** is in follow(B)

2 is in entry because **(** is in first(B**(**B**)**)

Table for LL(1) Parsing

Tells which alternative to take, given current token:

choice : Nonterminal x Token -> Set[Int]

A ::= (1) $B_1 \ldots B_p$
| (2) $C_1 \ldots C_q$
| (3) $D_1 \ldots D_r$

if $t \in$ first($C_1 \ldots C_q$) add 2 to choice(A,t)

if $t \in$ follow(A) add K to choice(A,t) where K is nullable

For example, when parsing A and seeing token t

choice(A,t) = {2} means: parse alternative 2 ($C_1 \ldots C_q$)

choice(A,t) = {3} means: parse alternative 3 ($D_1 \ldots D_r$)

choice(A,t) = {} means: report syntax error

choice(A,t) = {2,3} : not LL(1) grammar

General Idea when parsing nullable(A)

$$A ::= B_1 \ldots B_p \mid C_1 \ldots C_q \mid D_1 \ldots D_r$$

def A =
  if (token ∈ T1) {
    $B_1 \ldots B_p$
  else if (token ∈ **(**T2 U T_F)) {
    $C_1 \ldots C_q$
  } **else if** (token ∈ T3) {
    $D_1 \ldots D_r$
  } // no else error, just return

where:

$T1 =$ **first**$(B_1 \ldots B_p)$

$T2 =$ **first**$(C_1 \ldots C_q)$

$T3 =$ **first**$(D_1 \ldots D_r)$

$T_F =$ **follow**(A)

Only one of the alternatives can be nullable (here: 2nd)

T1, T2, T3, T_F should be pairwise **disjoint** sets of tokens.