

# COMP4403 Crib Sheet

## Parsing Theory

### Context Free Grammars

Basic Example of Context Free Grammar

$E \rightarrow E \text{ Op } E$

$E \rightarrow "(E)"$

$E \rightarrow \text{number}$

$\text{Op} \rightarrow "+"$

$\text{Op} \rightarrow "-"$

$\text{Op} \rightarrow "*"$

Has start symbol  $E$ , nonterminals  $\{E, \text{Op}\}$ , and terminals

$\{(" , \backslash )", \text{number}, \backslash +, \backslash -, \backslash *, \backslash "\}$

A context-free grammar consists of:

- A finite set,  $\Sigma$ , of terminal symbols.
- A finite nonempty set of nonterminal symbols (disjoint from the terminal symbols).
- A finite nonempty set of productions of the form  $A \rightarrow \alpha$ , where  $A$  is a nonterminal symbol, and  $\alpha$  is a possibly empty sequence of symbols, each of which is either a terminal or nonterminal symbol.
- A start symbol that must be a nonterminal symbol

### Directly Derives

If there is a production in the form of  $N \rightarrow \gamma$  then we can directly derive  $\alpha N \beta \rightarrow \alpha \gamma \beta$ , where  $\alpha$  and  $\beta$  are possibly empty sequences of terminal and nonterminal symbols.

### Derives

Given a sequence of terminal and nonterminal symbols,  $\alpha$ , derives a sequence  $\beta$ , written  $\alpha \xRightarrow{*} \beta$  if there is a finite sequence of zero or more direct derivation steps that start from  $\alpha$  and finishing with  $\beta$ , there must be one or more sequence  $\gamma_1, \gamma_2, \dots, \gamma_n$  such that  $\alpha = \gamma_0 \Rightarrow \gamma_1 \Rightarrow \dots \Rightarrow \gamma_n = \beta$ . Note that zero steps are allowed.

### Nullable

A possibly empty sequence of symbols,  $\alpha$ , is nullable if  $\alpha \xRightarrow{*} \epsilon$  or  $\alpha \xRightarrow{*} \text{Nullable rules}$

- *is nullable*
- any terminal symbol is not nullable
- a sequence of symbols is nullable if all of its constructs are nullable
- a set of alternatives is nullable if any of its constructs are nullable
- EBNF constructs for optionals and repetitions are nullable
- a nonterminal is nullable if there is a production with a nullable right-hand side

### Language

The formal language  $\mathcal{L}(G)$  corresponding to a Grammar,  $G$ , is:

$\mathcal{L}(G) = \{t \in \text{seq} \mid S \xRightarrow{*} t\}$

where  $S$  is the start symbol of  $G$  and  $\Sigma$  is its set of terminal symbols.

### Sentences and Sentential Form

A sequence of terminal symbols  $t$  such that

$S \xRightarrow{*} t$  is called a *sentence* of the language.

### Ambiguous grammars for sequence

A grammar,  $G$ , is ambiguous for a sentence,  $t$ , in  $\mathcal{L}$ , if there is more than one parse tree for  $t$ .

### Ambiguous grammars

A grammar,  $G$ , is ambiguous if it is ambiguous for any sentence in  $\mathcal{L}(G)$ .

### Left and right associative operators

To remove the ambiguity and treat  $"-"$  as a left associate operator (as usual) we can rewrite the grammar to

$E \rightarrow E "-" T$

$E \rightarrow T$

$T \rightarrow N$

and to treat  $"-"$  as right associative we use

$E \rightarrow T "-" E$

$E \rightarrow T$

$T \rightarrow N$

### Dangling Else

Dangling else is a problem that arises when we have an if-then-else statement and the else is ambiguous as to which if it belongs to.

## Recursive-Descent Parsing

### Backus-Naur Form (BNF)

Allows a context-free grammar to be described by a set of productions of the form  $N \rightarrow \alpha$  Where  $N$  is a nonterminal symbol and  $\alpha$  is a (possibly empty) sequence of terminal and nonterminal symbols  
The set of productions:

$N \rightarrow \alpha_1$

$N \rightarrow \alpha_2$

$\dots$

$N \rightarrow \alpha_n$

can be written as  $N \rightarrow \alpha_1 | \alpha_2 | \dots | \alpha_n$

### Extended BNF (EBNF)

Extends BNF to allow: an *optional* syntactic construct [in square brackets]

$\text{RelCondition} \rightarrow \text{Exp} [\text{RelOp Exp}]$  A

*repetition* construct in curly braces

$\text{StatementList} \rightarrow$

$\text{Statement} \{ \text{SEMICOLON Statement} \}$

The extended notation does not add any expressive power - any EBNF grammar can be converted to a BNF grammar...

### Translating EBNF to BNF

Replace optional construct  $[S]$  by a new nonterminal  $\text{Opt}S$ :

$\text{Opt}S \rightarrow S | \epsilon$

For example, we can rewrite

$\text{RelCondition} \rightarrow \text{Exp} [\text{RelOp Exp}]$  as

$\text{RelCondition} \rightarrow \text{Exp OptRelExp}$

$\text{OptRelExp} \rightarrow \text{RelOp Exp} | \epsilon$

Replace grouping construct  $( S )$  by a new nonterminal  $\text{Grp}S$ :  $\text{Grp}S \rightarrow S$

For example, we can rewrite  $\text{Rep}F \rightarrow$

$(\text{TIMES} | \text{DIVIDE}) \text{Factor Rep}F | \epsilon$

as

$\text{Rep}F \rightarrow \text{TermOp Factor Rep}F | \epsilon$

$\text{TermOp} \rightarrow \text{TIMES} | \text{DIVIDE}$

### Recursive-descent parsing

A recursive-descent parser is a recursive program to recognise sentences in a language:

- Input is a stream of lexical tokens, represented by tokens of type `TokenStream`.
- Each nonterminal symbol  $N$  has a method *parseN* that:
  - recognises the longest string of lexical tokens in the input stream, starting from the current token, derivable from  $N$ ;
  - as it parses the input it moves the current location forward;
  - when it has finished parsing  $N$ , the current token is the token immediately following the last token matched as part of  $N$  (which may be at the end-of-file token).

### matching a list of alternatives: example

$\text{Factor} \rightarrow$

$\text{LPAREN RelCondition RPAREN} |$

$\text{NUMBER} |$

$\text{LValue}$

```
void parseFactor () {
    if (tokens.isMatch(Token.LPAREN)) {
        tokens.match(Token.LPAREN);
        parseRelCondition();
        tokens.match(Token.RPAREN);
    } else if (tokens.isMatch(Token.NUMBER)) {
        tokens.match(Token.NUMBER);
    } else if (tokens.isMatch(Token.IDENTIFIER)) {
        parseLValue();
    } else {
        errors.error("Syntax error");
    }
}
```

}

### matching an optional construct: example

$\text{RelCondition} \rightarrow \text{Exp} [\text{RelOp Exp}]$

```
void parseRelCondition () {
    parseExp();
    if (tokens.isIn(REL_OPS_SET)) {
        parseRelOp();
        parseExp();
    }
}
```

### matching a repetition construct: example

$\text{Term} \rightarrow \text{Factor} (\text{TIMES} | \text{DIVIDE}) \text{Factor}$

```
void parseTerm () {
    parseFactor();
    while (tokens.isIn(TERM_OPS_SET)) {
        if (tokens.isMatch(Token.TIMES)) {
            tokens.match(Token.TIMES);
        } else if (tokens.isMatch(Token.DIVIDE)) {
            tokens.match(Token.DIVIDE);
        } else {
            fatal("Internal Error");
            // unreachable
        }
        parseFactor();
    }
}
```

### recovery strategy for matching a single token

```
void parseWhileStatement () {
    tokens.match(Token.KW_WHILE);
    parseCondition();
    tokens.match(Token.KW_DO,
        STATEMENT_START_SET);
    parseStatement();
}
```

Generally, it is

`tokens.match(expected, follows)`

maybe add parsing recovery sets here?

### LL(1) grammars

We indicated above that there are some restrictions on grammars to ensure they are suitable for recursive-descent predictive parsing. The class of grammars that are suitable is referred to as LL(1), where the first "L" refers to the fact that the parsing of the input is from Left to right, the second "L" refers to the fact that they produce a Leftmost derivation sequence, and the "1" indicates that their parsers use one symbol lookahead (i.e., the current token is the lookahead).

**Definition 1 (LL(1) Grammar)** A BNF grammar is LL(1) if for each nonterminal,  $N$ , the first sets for each pair of alternative productions for  $N$  are disjoint, and if  $N$  is nullable,  $\text{First}(N)$  and  $\text{Follow}(N)$  are disjoint

## First & Follow Sets

### First Sets

The first set for a construct  $\alpha$  records

- the set of terminal symbols  $\alpha$  can start with, and
- if  $\alpha$  is nullable, it contains the empty string  $\epsilon$  to indicate that.

### Calculating First Sets

Let

- $\alpha$  be a terminal symbol,
- $\alpha_1, \alpha_2, \dots, \alpha_n$  be strings of (terminal and nonterminal) symbols, and
- $A$  be a nonterminal symbol defined by a single production

$$A \rightarrow \alpha_1 | \alpha_2 | \dots | \alpha_n,$$

then

$$\text{Fst}(\epsilon) = \{\epsilon\}$$

$$\text{Fst}(\alpha) = \{\alpha\}$$

$$\text{Fst}(\alpha_1 | \alpha_2 | \dots | \alpha_n) = \text{Fst}(\alpha_1) \cup \text{Fst}(\alpha_2) \cup \dots \cup \text{Fst}(\alpha_n)$$

$$\text{Fst}(A) = \text{Fst}(\alpha_1 | \alpha_2 | \dots | \alpha_n)$$

Let  $S_1, S_2, \dots, S_n$  be (terminal or nonterminal) symbols, then

$$\text{Fst}(S_1 S_2 \dots S_n) =$$

$$\text{Fst}(S_1) - \{\epsilon\}$$

$$\cup \text{Fst}(S_2) - \{\epsilon\} \quad \text{if } S_1 \text{ is nullable}$$

$$\cup \dots$$

$$\cup \text{Fst}(S_i) - \{\epsilon\} \quad \text{if } S_1 - S_{i-1} \text{ is nullable}$$

$$\cup \dots$$

$$\cup \text{Fst}(S_n) - \{\epsilon\} \quad \text{if } S_1 - S_{n-1} \text{ is nullable}$$

$$\cup \{\epsilon\} \quad \text{if } S_1 - S_n \text{ is nullable}$$

### Calculating Algorithmically

Start with the first sets for all nonterminals being the empty set and note that the first set for every terminal symbol,  $\alpha$ , is the singleton set  $\{\alpha\}$ . We then make a pass over all production in a grammar considering all alternatives and process as follows.

- If there is a production of the form  $N \leftarrow \epsilon$ , we add  $\epsilon$  to the first set of  $N$ .

- If there is a production of the form  $N \leftarrow S_1 S_2 \dots S_n$ , then for each  $i \in 1..n$ , if for all  $j \in 1..i-1$ ,  $S_j$  is nullable, we add the current first set for  $S_i$  minus  $\epsilon$  to the first set for  $N$ .
- If every construct  $S_1, \dots, S_n$  is nullable, we add  $\epsilon$  to the first set for  $N$ .

We repeat the passes until no set is modified in the pass, in which case we are finished.

*Example*

$$A \leftarrow B x | C \quad (1)$$

$$B \leftarrow C y | D \quad (2)$$

$$C \leftarrow D z | \epsilon \quad (3)$$

$$D \leftarrow A w \quad (4)$$

$A$	$\{\}$	$\{\epsilon\}$	$\{\epsilon, y\}$	$\{\epsilon, y, w\}$	$\{\epsilon, y, w\}$
$B$	$\{\}$	$\{y\}$	$\{y\}$	$\{y, w\}$	$\{y, w\}$
$C$	$\{\epsilon\}$	$\{\epsilon\}$	$\{\epsilon, w\}$	$\{\epsilon, w, y\}$	$\{\epsilon, w, y\}$
$D$	$\{\}$	$\{w\}$	$\{w, y\}$	$\{w, y\}$	$\{w, y\}$

### Follow Sets

The follow set for a nonterminal,  $N$ , is the set of terminal symbols that may follow  $N$  in any context within the grammar. End-of-file is represented by the special terminal symbol  $\$$  in which  $a$  follows  $N$ .

A nonterminal,  $N$ , is followed by a terminal symbol,  $a$ , if there is a derivation from  $S\$$

### Calculating Follow Sets

We compute the Follow set for a nonterminal,  $N$ , using two rules.

- If there is a production of the form

$$A \rightarrow \alpha N \beta$$

then any symbols that can start  $\beta$  can follow  $N$ , and hence  $\text{Follow}(N)$  must include all the terminal symbols in  $\text{First}(\beta)$ . Note that  $\epsilon$  is not included even if it appears in  $\text{First}(\beta)$ .

$$\text{First}(\beta) - \{\epsilon\} \subseteq \text{Follow}(N)$$

- If there is a production of the form

$$A \rightarrow \alpha N \beta$$

and  $\beta$  is nullable, then any token that can follow  $A$  can also follow  $N$ . Hence,

$$\text{Follow}(A) \subseteq \text{Follow}(N)$$

The case where  $\beta$  is nullable includes the case when  $\beta$  is empty and the production is of the form

$$A \rightarrow \alpha N$$

### Calculating Algorithmically

Start with all nonterminal symbols having an empty follow set,  $\{\}$ , except for the start symbol,  $S$ , which has the follow set  $\{\$ \}$ .

We make a pass through the grammar examining the right side of every production. For each occurrence of a nonterminal within the right side of some production, we augment the Follow set for that nonterminal according to the following process. Assume we are processing an occurrence of  $N$  and the production is of the form

$$A \rightarrow \alpha N \beta$$

we add  $\text{First}(\beta) - \{\epsilon\}$  to the Follow set computer for  $N$  so far, and if  $\beta$  is nullable, we also add the current Follow set for  $A$  to the Follow set for  $N$ . After making a complete pass, we repeat the process with the Follow sets computed so far until no Follow sets are modified, in which case we are done.

*Example*

$$S \rightarrow xAB \quad \text{Fst}(S) = \{x\}$$

$$A \rightarrow y|zB \quad \text{Fst}(A) = \{y, z\}$$

$$B \rightarrow \epsilon | Ax \quad \text{Fst}(B) = \{\epsilon, y, z\}$$

$S$	$\{\$ \}$	$\{\$ \}$	$\{\$ \}$	$\{\$ \}$
$A$	$\{\}$	$\{y, z, \$, x\}$	$\{y, z, \$, x\}$	$\{y, z, \$, x\}$
$B$	$\{\}$	$\{\$, y, z\}$	$\{\$, y, z, x\}$	$\{\$, y, z, x\}$

## LL(1) Grammar

A BNF grammar is LL(1) if for each nonterminal,  $N$ , where  $N \rightarrow \alpha_1 | \alpha_2 | \dots | \alpha_n$ ,

- the First sets for each pair of alternatives for  $N$  are disjoint, and
- if  $N$  is nullable,  $\text{First}(N)$  and  $\text{Follow}(N)$  are disjoint.

## Left Factoring & Left Recursion

### Left Factoring Productions

Not all EBNF grammars are suitable for Recursive-Descent Parsing, however, sometimes we can rewrite them into a form that is suitable.

### Left Factor Rewriting Rule

To remove the left factor from

$$A \rightarrow \alpha \beta | \alpha \gamma,$$

we can rewrite the production using an additional nonterminal  $A'$  as

$$A \rightarrow \alpha A'$$

$$A' \rightarrow \beta | \gamma$$

### Left Recursive Productions

A production of the form

$$E \rightarrow E + T | T$$

is not suitable for RDP because the left recursion in the grammar leads to an infinite recursion.

### Immediate Left Recursion Rewriting Rule (Simple)

To remove the left recursion from

$$A \rightarrow A \alpha | \beta,$$

we can rewrite the production as

$$A \rightarrow \beta A'$$

$$A' \rightarrow \epsilon | \alpha A'$$

### Immediate Left Recursion Rewriting Rule (General)

To remove the left recursion from the general case

$$A \rightarrow A \alpha_1 | A \alpha_2 | \dots | A \alpha_n | \beta_1 | \beta_2 | \dots | \beta_m,$$

we can use grouping to see the structure in the same form as the simple case

$$A \rightarrow A(\alpha_1 | \alpha_2 | \dots | \alpha_n) | (\beta_1 | \beta_2 | \dots | \beta_m)$$

which allows us to rewrite the production as follows,

$$A \rightarrow (\beta_1 | \beta_2 | \dots | \beta_m) A'$$

$$A' \rightarrow \epsilon | (\alpha_1 | \alpha_2 | \dots | \alpha_n) A'$$

### Indirect Left Recursion Rewriting Rule

The following productions have indirect recursion from  $A \rightarrow B \rightarrow C \rightarrow A$ .

$$A \rightarrow B \alpha$$

$$B \rightarrow C \beta$$

$$C \rightarrow A \gamma_1 | \gamma_2$$

To remove the recursion we can first collapse  $B$  into  $A$  as follows.

$$A \rightarrow C \beta \alpha$$

$$C \rightarrow A \gamma_1 | \gamma_2$$

Then we can collapse  $C$  into  $A$ .

$$A \rightarrow (A \gamma_1 | \gamma_2) \beta \alpha$$

$$\downarrow$$

$$A \rightarrow A \gamma_1 \beta \alpha | \gamma_2 \beta \alpha$$

This now leaves a simple direct left recursion which we can remove as follows.

$$A \rightarrow \gamma_2 \beta \alpha A'$$

$$A' \rightarrow \gamma_1 \beta \alpha A'$$