Compiladores (IF688)

Leopoldo Teixeira

Imt@cin.ufpe.br | @leopoldomt

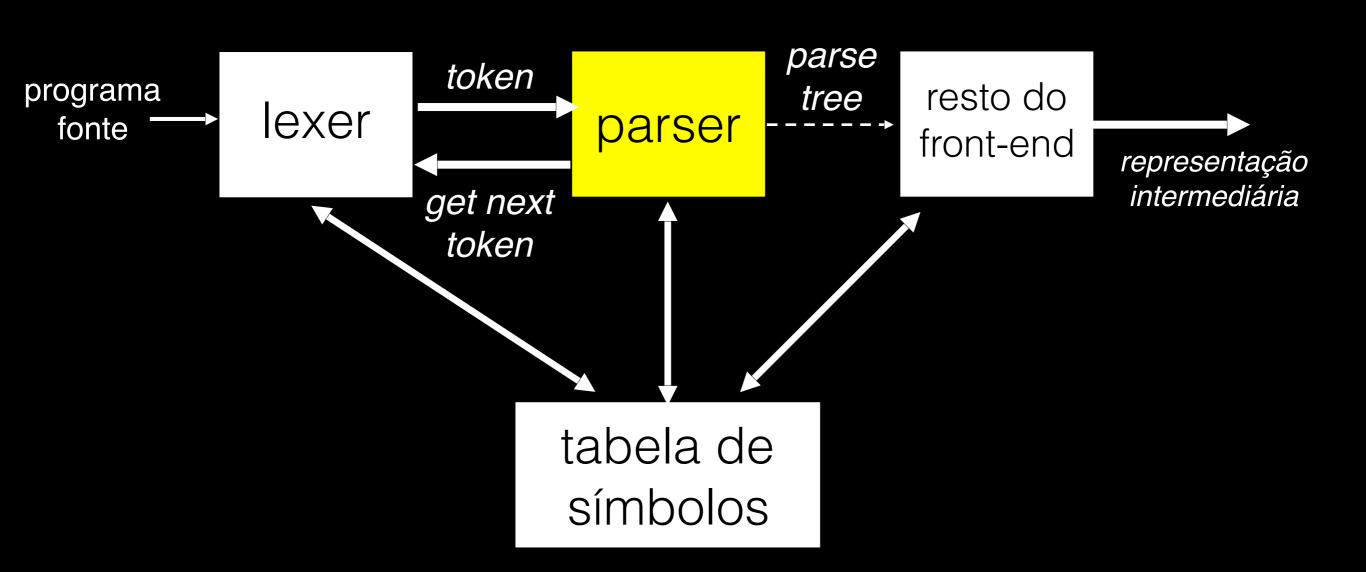
Contato

 Grupo de Discussão: https://classroom.google.com

Faça o login com sua conta do Cln. Na página inicial, clique em +. Digite na caixa o código da turma: **mpg9k1** Clique em PARTICIPAR

• E-mail: Imt@cin.ufpe.br Sala C012

Analisador Sintático



Parsing

→ parser → ou

- Caso sucesso:
 - A sintaxe do programa está correta
 - A string de entrada é "bem formada"
- Caso contrário:
 - erro de sintaxe; alguma regra sintática foi violada
- Importante:
 - O programa pode ainda conter erros capturados ou não pelo type checker

Gramáticas

- Especificações precisas e fáceis de entender da sintaxe de uma linguagem de programação
- Para algumas classes de gramáticas, é possível gerar automaticamente um parser eficiente.
- Tem relação direta com a estrutura da linguagem usada para tradução/compilação.
- Fáceis de serem estendidas e atualizadas.

Gramáticas Livres de Contexto

- $stmt \rightarrow if expr then stmt else stmt$
- Um conjunto de símbolos terminais (tokens).
- Um conjunto de símbolos não-terminais.
- Um conjunto de produções, cada produção consiste de um não-terminal, uma seta, e uma seqüencia de tokens e/ou não terminais
- Um não terminal designado como símbolo inicial

Expressões Regulares **vs.** Gramáticas Livres de Contexto

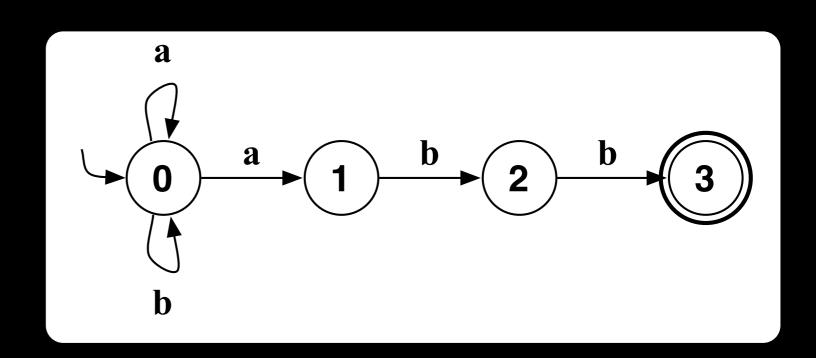
- Tudo que pode ser descrito por uma ER pode ser descrito com GLC, mas:
 - Regras léxicas são especificadas mais simplesmente com ER
 - ER geralmente são mais concisas e simples
 - Podem ser gerados analisadores léxicos mais eficientes a partir de expressões regulares
 - Estrutura/modulariza o front-end do compilador

(a b)*abb

$$A_0 \longrightarrow \mathbf{a}A_0 \mid \mathbf{b}A_0 \mid \mathbf{a}A_1$$
 $A_1 \longrightarrow \mathbf{b}A_2$

$$A_2 \rightarrow bA_3$$

$$A_3 \rightarrow \epsilon$$



Expressões Regulares **vs.**Gramáticas Livres de Contexto

- Expressões regulares são convenientes para especificar a estrutura de construções léxicas, como identificadores, constantes, palavras chave etc.
- Usamos gramáticas para especificar estruturas aninhadas, como parênteses, begin-end, if-thenelse etc.

Limites de especificação com gramáticas livres de contexto

- L = {wcw | w está em (a|b)*}
 ex: declaração de variáveis antes de seu uso
- L = {aⁿb^mcⁿd^m | n >= 1 e m >= 1}
 ex: contagem do número de argumentos de funções/procedimentos.

Que palavras pertencem a L(G)?

- aba
- acca
- abcba
- abcbcba

$$S \longrightarrow aXa$$

 $X \longrightarrow bY \mid \varepsilon$
 $Y \longrightarrow cXc \mid \varepsilon$

Derivação

 Derivação: Dada uma gramática G, produz uma string s que faz parte de L(G)

Gramática G

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

Derivação de abbebede

S

 $\rightarrow a\underline{A}Be$

 \rightarrow aAbcBe

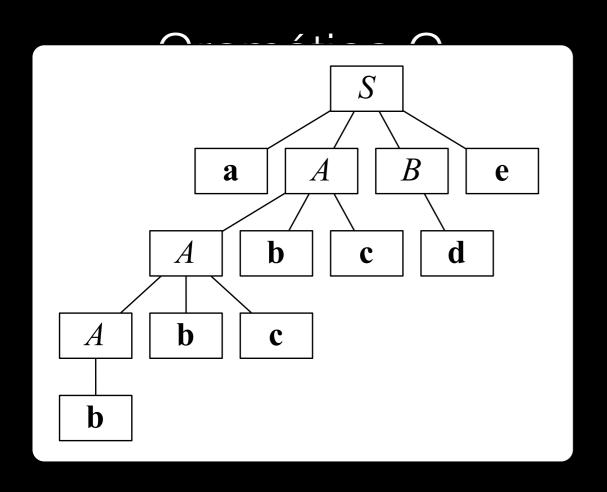
 \rightarrow a<u>A</u>bcbcBe

 \rightarrow abbcbc<u>B</u>e

→ abbcbcde

Parsing

 Dada uma string s em L(G), produz uma árvore sintática que demonstra como obter a derivação de s



Derivação de abbcbcde

S

- $\rightarrow a\underline{A}Be$
- $\rightarrow a\underline{A}bcBe$
- $\rightarrow a\underline{A}bcbcBe$
- \rightarrow abbcbc<u>B</u>e
- → abbcbcde

Categorias

- Métodos de parsing universais: funcionam para qualquer gramática, mas são muito ineficientes (inviáveis para uso prático).
- Top down: constroem as *parse trees* a partir da raiz em direção às folhas.
- Bottom-up: constroem as parse trees a partir das folhas.

Top-Down Parsing

O Método

- A partir do símbolo inicial da gramática, consuma tokens da esquerda para direita
- Decida que produção aplicar, de acordo com o token retornado
- Continue até que
 - todas as folhas sejam símbolos terminais e não há mais tokens a serem lidos da entrada
 - há uma falta de correspondência entre a entrada e as folhas da parse tree parcialmente construída

No primeiro caso, tudo ok!

Eno segundo?

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

abbcbcde S

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$

$$A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$$

$$B \longrightarrow \mathbf{d}$$

abbcbcde S

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$

$$A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$$

$$B \longrightarrow \mathbf{d}$$

a b b c b c d e aABe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e ABe

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$

$$A \rightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$$

$$B \rightarrow \mathbf{d}$$

a b b c b c d e ABe

escolher uma produção!

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e ABe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e bBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e Be



restaura estado anterior à última escolha

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$

$$A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$$

$$B \longrightarrow \mathbf{d}$$

faz outra escolha!

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e ABe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e AbcBe

Símbolo A novamente na produção mais à esquerda

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$

$$A \rightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$$

$$B \rightarrow \mathbf{d}$$

De novo! Escolher uma produção!

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e AbcBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e bbcBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e bcBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e cBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

abbcbcdeBe



restaura estado anterior à última escolha

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$

$$A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$$

$$B \longrightarrow \mathbf{d}$$

faz outra escolha!

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e AbcBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e AbcbcBe

Símbolo A novamente na produção mais à esquerda

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$

$$A \rightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$$

$$B \rightarrow \mathbf{d}$$

De novo! Escolher uma produção!

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e AbcbcBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e bbcbcBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e bcbcBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e cbcBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e bcBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e cBe

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e Be

Símbolo B na produção mais à esquerda

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$

$$A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$$

$$B \longrightarrow \mathbf{d}$$

Só tem uma escolha...

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e Be

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e de

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$

a b b c b c d e e

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$
 $A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$
 $B \longrightarrow \mathbf{d}$



Fim do arquivo

- Parsers devem ler não apenas os símbolos terminais, mas também o marcador de fim de arquivo
- Usamos \$ para representar o fim do arquivo
 - $S' \rightarrow SS$

$$S \longrightarrow \mathbf{a}AB\mathbf{e}$$

$$A \longrightarrow A\mathbf{b}\mathbf{c} \mid \mathbf{b}$$

$$B \longrightarrow \mathbf{d}$$

a b b c b c d e \$

Derivação correspondente!

S

 $\rightarrow a\underline{A}Be$

 $\rightarrow a\underline{A}bcBe$

 \rightarrow aAbcbcBe

 \rightarrow abbcbc<u>B</u>e

→ abbcbcde

Backtracking: A maior fonte de ineficiência em um parser top-down e leftmost

Seria interessante termos parsers que não fazem backtracking

predictive parsing

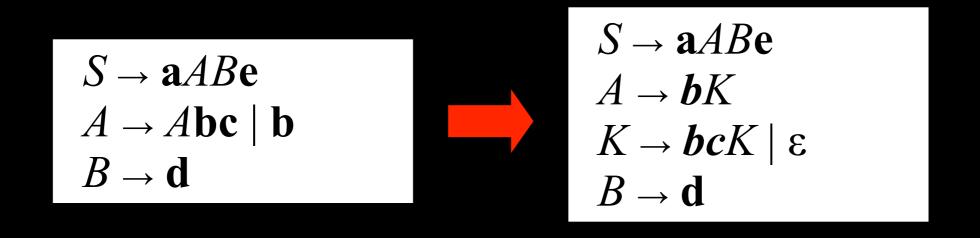
O que é importante na gramática para que o parser seja preditivo?

O token lido como primeiro terminal deve fornecer informação suficiente para decidirmos que produção aplicar

Que situações normalmente geram problemas para predictive parsing?

Recursão à esquerda

- Uma gramática é recursiva à esquerda se existe um não terminal A tal que existe uma derivação de A que gera $A\alpha$, para alguma string α .
- Existem técnicas para eliminar a recursão à esquerda, já vistas nas aulas anteriores.



Eliminando recursão à esquerda

$$S \rightarrow Aa \mid b$$
 $A \rightarrow Ac \mid Sd \mid \epsilon$

Eliminando recursão à esquerda

$$S \rightarrow Aa \mid b$$
 $A \rightarrow Ac \mid Aad \mid bd \mid \epsilon$

Solução

Fatoração à Esquerda

- Técnica de transformação de gramática usada para produzir uma gramática adequada para predictive parsing.
- Combina os casos em que há mais de uma alternativa possível a partir do reconhecimento de um token.
- Existem algoritmos para fazer a fatoração à esquerda.

Fatoração à Esquerda - exemplo

```
stmt \rightarrow if expr then stmt
stmt \rightarrow if expr then stmt else stmt
stmt \rightarrow other
```

Fatoração à Esquerda - exemplo

$$S \rightarrow iEtS|iEtSeS|a$$

 $E \rightarrow b$

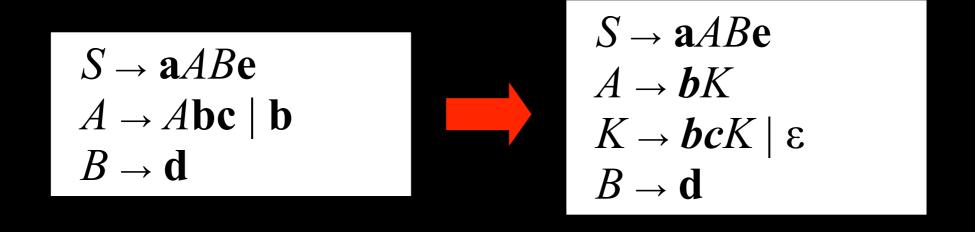
Fatoração à Esquerda - solução

$$S \rightarrow i E t S | i E t S e S | a$$
 $E \rightarrow b$

$$S \rightarrow i E t S S' | a$$
 $S' \rightarrow e S | \varepsilon$
 $E \rightarrow b$

Predictive parsing

- Na gramática original, A não tem informação suficiente para decidir qual produção aplicar ao enxergar um b no lookahead
- Na gramática modificada, cada produção alternativa do mesmo não terminal inicia com terminais distintos



Classificação de Parsers

Número de *tokens* de *lookahead* lidos, porém, não consumidos

Lê entrada da esquerda (L) para direita

LĻ(k)

Procura derivação mais à esquerda

FIRST e FOLLOW

- A construção de parsers top-down e bottom-up é auxiliada por duas funções, FIRST e FOLLOW
- Estas funções auxiliam o parser a decidir qual produção será aplicada, baseado no próximo símbolo de entrada
- Em condições de recuperação de erro, também podem ser úteis

FIRST

- FIRST(α), onde α é qualquer string de símbolos da gramática, é o conjunto de terminais que iniciam strings derivadas a partir de α
- Se α pode gerar ϵ , ϵ pertence a FIRST(α)
- Considere duas produções $A \rightarrow \alpha \mid \beta$
 - Para que o parser seja preditivo, o que deve ser verdade?

FIRST

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$
 $A \rightarrow \mathbf{b}K$
 $K \rightarrow \mathbf{b}\mathbf{c}K$
 $K \rightarrow \mathbf{c}$
 $B \rightarrow \mathbf{d}$

	FIRST	FOLLOW
S		
A		
K		
B		

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$
 $A \rightarrow \mathbf{b}K$
 $K \rightarrow \mathbf{b}\mathbf{c}K$
 $K \rightarrow \mathbf{c}$
 $B \rightarrow \mathbf{d}$

	FIRST	FOLLOW
S	{ a }	
\boldsymbol{A}		
K		
B		

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$

$$A \rightarrow \mathbf{b}K$$

$$K \rightarrow \mathbf{b}\mathbf{c}K$$

$$K \rightarrow \mathbf{c}$$

$$B \rightarrow \mathbf{d}$$

	FIRST	FOLLOW
S	{ a }	
\boldsymbol{A}	{ b }	
K		
B		

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$

$$A \rightarrow \mathbf{b}K$$

$$K \rightarrow \mathbf{b}\mathbf{c}K$$

$$K \rightarrow \mathbf{c}$$

$$B \rightarrow \mathbf{d}$$

	FIRST	FOLLOW
S	{ a }	
\boldsymbol{A}	{ b }	
K	{ b, ε }	
B		

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$

$$A \rightarrow \mathbf{b}K$$

$$K \rightarrow \mathbf{b}\mathbf{c}K$$

$$K \rightarrow \mathbf{c}$$

$$B \rightarrow \mathbf{d}$$

	FIRST	FOLLOW
S	{ a }	
\boldsymbol{A}	{ b }	
K	{ b, ε }	
B	{ d }	

Calculando FIRST(X)

- Se X é um terminal FIRST(X) = {X};
- Se $X \rightarrow \varepsilon$, $\varepsilon \in FIRST(X)$
- Se $X \rightarrow Y_1 Y_2 ... Y_k$, FIRST $(Y_1 Y_2 ... Y_k) \subseteq FIRST(X)$
- FIRST $(Y_1Y_2...Y_k)$ é
 - $[\varepsilon \notin FIRST(Y_1)] FIRST(Y_1)$; ou
 - $[\varepsilon \in FIRST(Y_1)] FIRST(Y_1) \{\varepsilon\} \cup FIRST(Y_2...Y_k)$
 - Se $\varepsilon \in FIRST(Y_i)$ para todo j de 1 a k, $\varepsilon \in FIRST(Y_1Y_2...Y_k)$

- FOLLOW(A), onde A é um não-terminal, é o conjunto de terminais α que pode aparecer imediatamente à direita de A
- Se A pode ser a última produção à direita, \$
 pertence a FOLLOW(A)

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$

$$A \rightarrow \mathbf{b}K$$

$$K \rightarrow \mathbf{b}\mathbf{c}K$$

$$K \rightarrow \mathbf{c}$$

$$B \rightarrow \mathbf{d}$$

	FIRST	FOLLOW
S	{ a }	
\boldsymbol{A}	{ b }	
K	{ b, ε }	
B	{ d }	

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$

$$A \rightarrow \mathbf{b}K$$

$$K \rightarrow \mathbf{b}\mathbf{c}K$$

$$K \rightarrow \mathbf{c}$$

$$B \rightarrow \mathbf{d}$$

	FIRST	FOLLOW
S	{ a }	{\$ }
A	{ b }	
K	{ b, ε }	
B	{ d }	

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$

$$A \rightarrow \mathbf{b}K$$

$$K \rightarrow \mathbf{b}\mathbf{c}K$$

$$K \rightarrow \mathbf{c}$$

$$B \rightarrow \mathbf{d}$$

	FIRST	FOLLOW
S	{ a }	{\$ }
\boldsymbol{A}	{ b }	{ d }
K	{ b, ε }	
B	{ d }	

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$

$$A \rightarrow \mathbf{b}K$$

$$K \rightarrow \mathbf{b}\mathbf{c}K$$

$$K \rightarrow \mathbf{c}$$

$$B \rightarrow \mathbf{d}$$

	FIRST	FOLLOW
S	{ a }	{\$ }
\boldsymbol{A}	{ b }	{ d }
K	{ b, ε }	{ d }
B	{ d }	

$$S \rightarrow \mathbf{a}AB\mathbf{e}$$

$$A \rightarrow \mathbf{b}K$$

$$K \rightarrow \mathbf{b}\mathbf{c}K$$

$$K \rightarrow \mathbf{c}$$

$$B \rightarrow \mathbf{d}$$

	FIRST	FOLLOW
S	{ a }	{\$ }
\boldsymbol{A}	{ b }	{ d }
K	{ b, ε }	{ d }
B	{ d }	{ e }

Calculando FOLLOW

- $\$ \in FOLLOW(S)$, onde S é o símbolo inicial e \$ é fim da entrada
- Se existe uma produção $A \to \alpha B\beta$, tudo que pertence a FIRST(β) exceto ϵ está em FOLLOW(B)
- Se existe uma produção $A \rightarrow \alpha B$, então tudo que estiver em FOLLOW(A) estará em FOLLOW(B)
- Se existe uma produção $A \to \alpha B\beta$, e $\epsilon \in FIRST(\beta)$, tudo que estiver em FOLLOW(A) estará em FOLLOW(B)