

Gramáticas Livres de Contexto

IBM1088 Linguagens Formais e Teoria da Computação

Evandro Eduardo Seron Ruiz
evandro@usp.br

UNIVERSIDADE DE SÃO PAULO

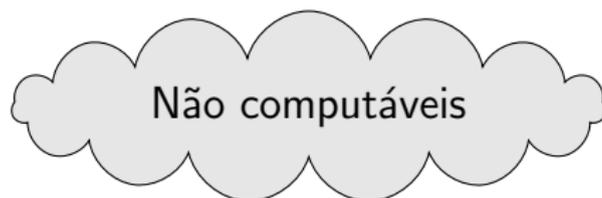
Quando vires um homem bom, tenta imitá-lo; quando vires um homem mau, examina-te a ti mesmo.

Confúcio^a

^aFilósofo chinês (551 A.C. — 479 A.C.).

- Até aqui estudamos as linguagens formais através de:
 - Autômatos finitos
 - Expressões Regulares
 - Gramáticas e linguagens regulares
- Vimos que linguagens simples como $L = \{a^n b^n \mid n \geq 0\}$ não podem ser descritas usando os conceitos acima
- Vamos estudar uma classe mais poderosa de linguagens, as **Linguagens livres de contexto**

Hierarquia de Chomsky



Recursivamente enumeráveis

Sensíveis ao contexto

Livres do contexto

Regulares

Gramáticas

Máquina de Turing

Autômato linearmente limitado

Autômato com pilha

Autômato de estado finito

Máquinas

- Definir formalmente as **Gramáticas livres de contexto**
- Mostrar como as GLC são uma evolução de LF em relação às GR
- Apresentar a Forma Normal de Chomsky

- 1 Introdução
- 2 Formalização de GLC
- 3 GCL ambíguas
- 4 Forma normal de Chomsky

Introdução: GLC

- Formalismo mais poderoso que as GR para descrever linguagens formais
- GLC descrevem as Linguagens Livres de Contexto (LLC)
- GLC permitem descrever o casamento de símbolos nas LF

Introdução: GLC

- Formalismo mais poderoso que as GR para descrever linguagens formais
- GLC descrevem as Linguagens Livres de Contexto (LLC)
- GLC permitem descrever o casamento de símbolos nas LF

Aplicações

- Especificação e compilação de linguagens de programação
- Analisador Léxico de LP
- Conhecimento poderoso na área de Processamento de Língua Natural
- Interface Homem-Máquina

GLC e ER

- GLC são essencialmente mais poderosas do que ER
- Qualquer linguagem que possa ser gerada por uma ER pode ser gerada por uma GLC
- Existem linguagens que podem ser geradas por GLC e não podem ser geradas por ER

GLC e ER

- GLC são essencialmente mais poderosas do que ER
- Qualquer linguagem que possa ser gerada por uma ER pode ser gerada por uma GLC
- Existem linguagens que podem ser geradas por GLC e não podem ser geradas por ER

Corolário

GLC são mais modelos mais poderosos que Autômatos Finitos.

Prova em duas partes:

- Dada uma ER R , podemos gerar uma GLC G tal que $L(R) = L(G)$
- Podemos definir G de tal modo que não exista AF F tal que $L(R) = L(G)$

Regras de produção

- Uma GLC é especificada através de **regras de produção** ou *regras de substituição* como, **por exemplo**:

$$V \rightarrow \alpha$$

- Para V é uma **variável**, ou ainda, um símbolo não-terminal,
- $\alpha \in (V \cup \Sigma)^*$, sendo Σ o conjunto dos símbolos terminais .

Veremos a definição completa mais adiante.

Vejam um exemplo de GLC a seguir...

Exemplo 1

Um exemplo de uma GLC $G_0 = (\{S\}, \{a, b\}, P_0, S)$ com

$$S \rightarrow aSb$$

$$S \rightarrow \epsilon$$

é tal que gera $L_0 = \{a^n b^n \mid n \geq 0\}$.

G_0 é uma quadrupla com P_0 sendo as regras de produção, S o símbolo inicial, além dos símbolos terminais e não terminais.

Exemplo 2

Neste exemplo vemos uma GLC, G_1 , com as seguintes regras de produção P_1 :

$$S \rightarrow AB$$

$$S \rightarrow ASB$$

$$A \rightarrow a$$

$$B \rightarrow b$$

Sendo que, S , A e B são **variáveis**, S é conhecido também como símbolo inicial. Ainda, a e b são símbolos **terminais**.

Exemplo 2

Neste exemplo vemos uma GLC, G_1 , com as seguintes regras de produção P_1 :

$$S \rightarrow AB$$

$$S \rightarrow ASB$$

$$A \rightarrow a$$

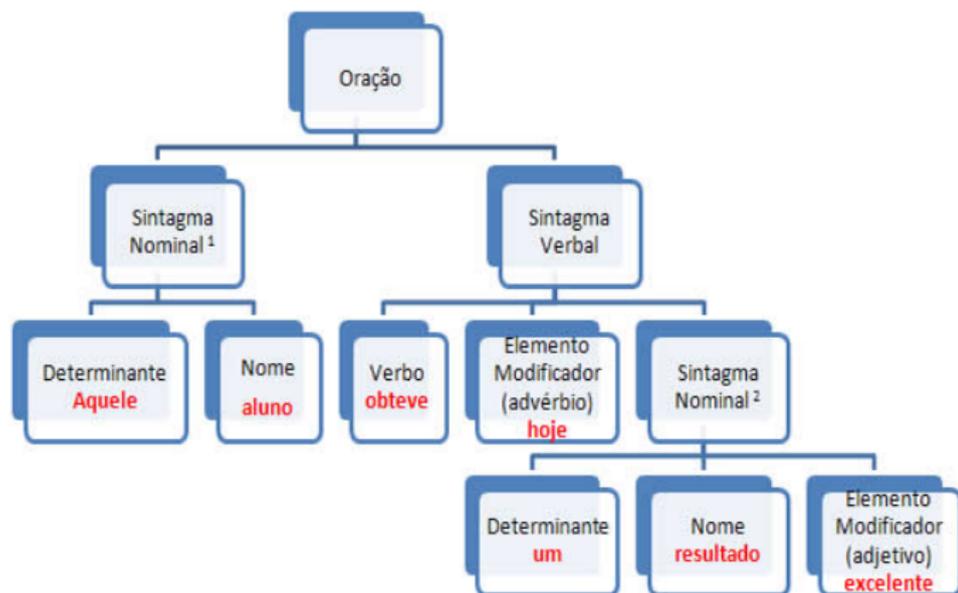
$$B \rightarrow b$$

Sendo que, S , A e B são **variáveis**, S é conhecido também como símbolo inicial. Ainda, a e b são símbolos **terminais**.

Pergunta

Quais cadeias G_1 produz e como ela se diferencia de G_0 ?

Sistema arbóreo para Análise Sintática



Fonte: www.portugues.com.br

Outro exemplo

Um outro exemplo interessante que descreve uma pequena parte da gramática da língua inglesa, é esta gramática abaixo:

$$S \rightarrow NP \quad VP$$
$$NP \rightarrow \text{det} \quad \text{noun}$$
$$VP \rightarrow \text{verb} \quad NP$$
$$\text{det} \rightarrow \text{a, the}$$
$$\text{noun} \rightarrow \text{boy, girl, book}$$
$$\text{verb} \rightarrow \text{took, saw, heard}$$

Veja explicação no próximo slide.

Outro exemplo: explicação

$$S \rightarrow NP \quad VP$$
$$NP \rightarrow \text{det} \quad \text{noun}$$
$$VP \rightarrow \text{verb} \quad NP$$
$$\text{det} \rightarrow \text{a, the}$$
$$\text{noun} \rightarrow \text{boy, girl, book}$$
$$\text{verb} \rightarrow \text{took, saw, heard}$$

Uma sentença S pode é formada por por um sintagma nominal (NP) seguido por um sintagma verbal (VP). O NP é composto de um artigo (det) e um substantivo (noun); enquanto o VP é constituído de um verbo (verb) e outro NP.

Outro exemplo: geração

Usando uma GLC podemos gerar

$$S \rightarrow NP \quad VP$$
$$NP \rightarrow \text{det} \quad \text{noun}$$
$$VP \rightarrow \text{verb} \quad NP$$
$$\text{det} \rightarrow \text{a, the}$$
$$\text{noun} \rightarrow \text{boy, girl, book}$$
$$\text{verb} \rightarrow \text{took, saw, heard}$$

Podemos gerar (ou analisar sintaticamente) sentenças como:

- the boy took the book
- the girl saw a boy
- the boy heard a dog

Definição: GLC

Definição de GLC

Uma **gramática livre de contexto** é uma 4-tupla (V, Σ, P, S) , em que:

- V é um conjunto finito de símbolos denominado **variáveis**,
- Σ é um conjunto finito de símbolos, disjunto de V , denominado **terminais**,
- P é um conjunto finito de **regras de produção** da forma $A \rightarrow \alpha$, tal que $A \in V$ e $\alpha \in (V \cup \Sigma)^*$.
- $S \in V$ é a variável inicial

Portanto, uma GLC é uma gramática na qual o lado esquerdo das produções contém exatamente uma variável.

Definição: LLC

Definição de LLC

Seja G uma **gramática livre de contexto**, $G = (V, \Sigma, P, S)$. A linguagem gerada por G ,

$$L(G) = \{w \in \Sigma^* \mid S \Longrightarrow^+ w\}$$

é dita uma **linguagem livre de contexto**, ou linguagem tipo 2.

Definição: LLC

Definição de LLC

Seja G uma **gramática livre de contexto**, $G = (V, \Sigma, P, S)$. A linguagem gerada por G ,

$$L(G) = \{w \in \Sigma^* \mid S \Longrightarrow^+ w\}$$

é dita uma **linguagem livre de contexto**, ou linguagem tipo 2.

O termo 'livre de contexto' vem do tipo de regra de produção que caracteriza a gramática. Como regra geral $A \rightarrow \alpha$, entende-se que a variável A deriva α sem depender de qualquer restrição ou análise dos símbolos que antecedem ou sucedem A .

Exemplo 1 (1/2)

Um exemplo de LLC é esta linguagem:

$$L_0 = \{a^n b^n \mid n \geq 0\},$$

já vista anteriormente, e que é originária da gramática:

$$G_0 = (\{S\}, \{a, b\}, P_1, S)$$

na qual as regras de produção são: $P_0 = \{S \rightarrow aSB \mid S \rightarrow \epsilon\}$.

Exemplo 1 (2/2)

L_1 gera, por exemplo, palavras com duplo balanceamento, tal como $aabb$. A seguinte derivação pode originar esta palavra:

$$S \implies aSB \implies aaSbb \implies aa\epsilon bb = aabb$$

Este tipo de linguagem é muito útil na análise sintática de linguagens de programação pois elas podem verificar, por exemplo, o balanceamento de parênteses, $(^n)^n$, e comandos de blocos estruturados, como os colchetes que são usados para delimitar estes blocos em linguagens como **C**, e Java.

Exemplo 2 (1/2)

Reparem que a linguagem L_1 gerada pela gramática G_1 a seguir é composta de expressões aritméticas contendo dois operadores '+' e '*', colchetes balanceados:

$$G_1 = (\{E\}, \{+, *, [,], x\}, P_1, E)$$

em que as regras de produção P_1 são:

$$E \rightarrow E + E$$

$$E \rightarrow E * E$$

$$E \rightarrow [E]$$

$$E \rightarrow x$$

Exemplo 2 (2/2)

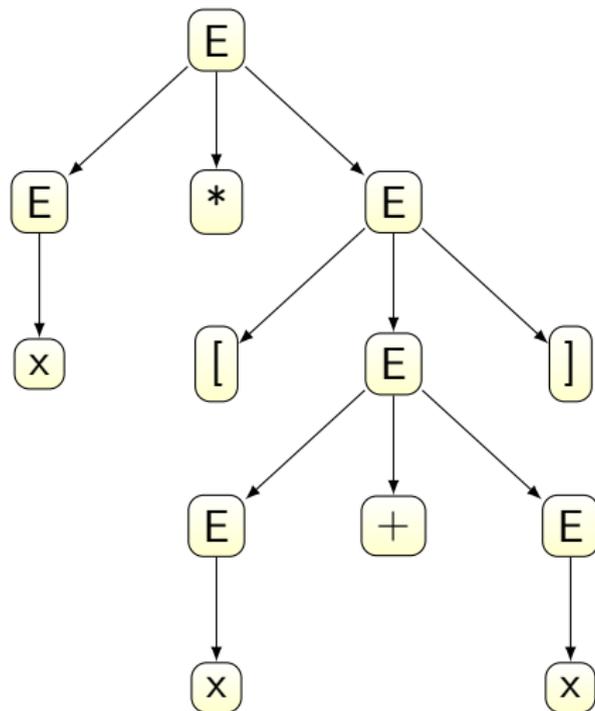
Deste modo, a expressão $x * [x + x]$ pode ser gerada pela seguinte derivação:

$$\begin{aligned} E &\Longrightarrow E * E \\ &\Longrightarrow x * E \\ &\Longrightarrow x * [E] \\ &\Longrightarrow x * [E + E] \\ &\Longrightarrow x * [x + E] \\ &\Longrightarrow x * [x + x] \end{aligned}$$

Árvore de Derivação

- Pode ser conveniente representar a derivação de palavras na forma de uma árvore
- **Árvore de Derivação**
- Raíz: variável inicial
- Folhas: símbolos terminais

Veja exemplo a seguir:

Árvore de derivação da expressão $x*[x+x]$ 

Exercício

Já!

Considerando G_1 descreva ao menos duas árvores de derivação para esta expressão: $x+x*x$.

Recordando as regras de produção P_1 são:

$$E \rightarrow E + E$$

$$E \rightarrow E * E$$

$$E \rightarrow [E]$$

$$E \rightarrow x$$

GCL ambíguas

- Pode ocorrer de uma palavra estar associada a mais de uma derivação
- Caracteriza **gramática ambígua**
- Propriedade não interessante para muitas aplicações
 - e.g.: Compiladores. Uma análise sintática pode gerar dois códigos distintos

GLC ambígua: definição

GLC ambígua: definição

Uma gramática livre do contexto é uma GCL ambígua se existe pelo menos uma palavra que possua duas ou mais derivações.

Vejamos quatro formas de derivação da palavra $x + x * x$ da gramática $G_1 = (\{E\}, \{+, *, [,], x\}, P_1, E)$, vista anteriormente.

Derivações mais à esquerda

Uma derivação

$$\begin{aligned} E &\Rightarrow E + E \\ &\Rightarrow x + E \\ &\Rightarrow x + E * E \\ &\Rightarrow x + x * E \\ &\Rightarrow x + x * x \end{aligned}$$

Derivações mais à esquerda

Uma derivação

$$\begin{aligned}
 E &\Rightarrow E + E \\
 &\Rightarrow x + E \\
 &\Rightarrow x + E * E \\
 &\Rightarrow x + x * E \\
 &\Rightarrow x + x * x
 \end{aligned}$$

Outra derivação

$$\begin{aligned}
 E &\Rightarrow E * E \\
 &\Rightarrow E + E * E \\
 &\Rightarrow x + E * E \\
 &\Rightarrow x + x * E \\
 &\Rightarrow x + x * x
 \end{aligned}$$

Derivações mais à direita

Uma derivação

$$\begin{aligned} E &\Rightarrow E + E \\ &\Rightarrow E + E * E \\ &\Rightarrow E + E * x \\ &\Rightarrow E + x * x \\ &\Rightarrow x + x * x \end{aligned}$$

Derivações mais à direita

Uma derivação

$$\begin{aligned}
 E &\Longrightarrow E + E \\
 &\Longrightarrow E + E * E \\
 &\Longrightarrow E + E * x \\
 &\Longrightarrow E + x * x \\
 &\Longrightarrow x + x * x
 \end{aligned}$$

Outra derivação

$$\begin{aligned}
 E &\Longrightarrow E * E \\
 &\Longrightarrow E * x \\
 &\Longrightarrow E + E * x \\
 &\Longrightarrow E + x * x \\
 &\Longrightarrow x + x * x
 \end{aligned}$$

Exercício

Observação!

Apesar das dificuldades impostas pela ambigüidade de algumas GLC, é importante salientar que algumas LLC só podem ser geradas por gramáticas ambíguas.

Exercício

Observação!

Apesar das dificuldades impostas pela ambigüidade de algumas GLC, é importante salientar que algumas LLC só podem ser geradas por gramáticas ambíguas.

Exercício. Já!

Mostre que a linguagem $L_2 = \{a^i b^j c^k \mid i = j \text{ ou } i = k\}$ é ambígua. Lembre-se que primeiro devemos construir uma gramática que expresse esta linguagem.

Resposta

$G = (V, \Sigma, P, S)$ para $V = S, W, X, Y, Z$, sendo S a variável inicial, e os terminais $\Sigma = \{a, b, c\}$. Veja as regras de produção:

$$S \rightarrow XY \mid W$$

$$X \rightarrow aXb \mid \epsilon$$

$$Y \rightarrow cY \mid \epsilon$$

$$W \rightarrow aWc \mid Z\epsilon$$

$$Z \rightarrow bZ \mid \epsilon$$

Simplificação de GLC ambígüa

Podemos simplificar uma GCC ambígüa sem reduzir seu poder de expressão usando as seguintes técnicas:

Símbolos inúteis Exclusão de variáveis ou terminais que não são usados para gerar palavras

Produções vazias Eliminação de produções tipo $A \rightarrow \epsilon$.¹

Substituição de variáveis Exclusão de produções como $A \rightarrow B$, ou seja, uma substituição de variável que não acrescenta poder de geração de palavras à gramática

¹Quando a palavra ϵ pertence a uma linguagem ela deve ser incluída usando-se uma regra de produção específica para este fim, geralmente $S \rightarrow \epsilon$.

Forma normais

As formas normais são consideradas como

- Técnica de descrição da gramáticas livres do contexto, ou melhor
- Regras de notação para descrição das regras de produção das GLC

Em geral, os diferentes tipos de formas normais estabelecem regras rígidas na descrição das regras de produção, como é o caso da Forma Normal de Chomsky (FNC) e da Forma Normal de Backus–Naur.

Forma normal de Chomsky

FNC

Uma GLC $G = (V, \Sigma, P, S)$ é dita estar na **Forma Normal de Chomsky** se todas as regras de produção estão escritas num dos seguintes modos:

$$A \rightarrow BC \quad \text{ou} \quad A \rightarrow a$$

para $A, B, C \in V$, $a \in \Sigma$, sendo que B e C não podem ser a variável inicial. Permite-se $S \rightarrow \epsilon$ para S como a variável inicial.

FNC e GLC

Teorema

Qualquer linguagem livre do contexto é gerada por uma GLC na Forma Normal de Chomsky.

Transformando uma GLC qualquer em Forma Normal de Chomsky, (1/2)

Etapa 1 Simplificação

Exclua as produções vazias, tipo $A \rightarrow \epsilon$. Exclua também as produções unitárias tipo $A \rightarrow B$, ou seja, de alteração de variável; como também os símbolos inúteis, ou não utilizados.

Etapa 2 Transformação do lado direito das produções de comprimento maior ou igual a dois

O objetivo desta etapa é garantir que o lado direito das produções de comprimento maior ou igual a dois seja composto exclusivamente por variáveis. A exclusão de um terminal pode acontecer desde que seja substituído por uma variável intermediária C_a e desde que incluída a produção $C_a \rightarrow a$.

Transformando uma GLC qualquer em Forma Normal de Chomsky, (2/2)

- Etapa 3** Transformação do lado direito das produções de comprimento maior ou igual a três em produções com exatamente duas variáveis
- O objetivo desta etapa é garantir que o lado direito das produções de comprimento maior ou igual a um seja composto exatamente por duas variáveis.

Exemplo

Considere a gramática livre do contexto G abaixo e simplifique suas produções para descrevê-la na Forma Normal de Chomsky. Seja

$$G = (\{E\}, \{+, *, [,], x\}, P, E)$$

para as produções P , como seguem:

$$E \rightarrow E + E$$

$$E \rightarrow E * E$$

$$E \rightarrow [E]$$

$$E \rightarrow x$$

Aplicando as regras em G para G_{FNC} , veja:

Aplicando as regras

Etapa 1 A gramática já esta de acordo.

Etapa 2 As 3 primeiras produções precisam ser modificadas usando-se variáveis intermediárias desta maneira:

$$E \rightarrow EC_+E$$

$$E \rightarrow EC_*E$$

$$E \rightarrow C_+[EC_]$$

$$C_+ \rightarrow +$$

$$C_* \rightarrow *$$

$$C_[] \rightarrow [$$

$$C_[] \rightarrow]$$

continuação...

Etapa 3 As três primeiras produções precisam ser substituídas para:

$$E \rightarrow ED_1$$

$$E \rightarrow ED_2$$

$$E \rightarrow C_1 D_3$$

$$D_1 \rightarrow C_+ E$$

$$D_2 \rightarrow C_* E$$

$$D_3 \rightarrow EC_1$$

Por último...

Finalizando, a gramática livre do contexto G fica assim:

$$G_{FNC} = (\{E, C_+, C_*, C[, C], D_1, D_2, D_3\}, \{+, *, [,], x\}, P', E)$$

para as produções P' , como seguem:

Regras de produção

$$E \rightarrow ED_1$$

$$E \rightarrow ED_2$$

$$E \rightarrow C_1 D_3$$

$$E \rightarrow x$$

$$D_1 \rightarrow C_+ E$$

$$D_2 \rightarrow C_* E$$

$$D_3 \rightarrow EC_1$$

$$C_+ \rightarrow +$$

$$C_* \rightarrow *$$

$$C_1 \rightarrow [$$

$$C_1 \rightarrow]$$

Caros,

Procurem fixar o conteúdo lendo os livros texto e fazendo exercícios.