Linguagens Formais

João Paulo Lazzarini Cyrino

06/10/2020

Teoria da Gramática

Podemos pensar em **Gramática** como um dispositivo que caracteriza todas as sentenças de uma língua. Uma Teoria da Gramática poderia se dar de três formas (segundo Chomsky, em *Logical Structure of Linguistic Theory*), com graus de dificuldade descrescentes:

1. **Dispositivo de descoberta (melhor teoria):** Um dispositivo que, alimentado pelos dados de uma língua L, forneça sua gramática G.
2. **Dispositivo de verificação (segunda melhor teoria):** Um dispositivo que, alimentado pelos dados de uma língua L e por uma gramática G, diga se G é uma gramática da língua L.
3. **Dispositivo de avaliação (teoria "reserva"):** Um dispositivo que, alimentado pelos dados de uma língua L e por duas gramáticas $G1$ e $G2$, diga se $G1$ é melhor que $G2$.

Para Chomsky, nem a primeira nem a segunda opções seriam possíveis (nos anos 50~60). Não seria possível obter um dispositivo de generalização a partir dos dados, mas sim um dispositivo que estabeleça critérios formais para qual a melhor análise para um conjunto de dados.

O porém dessa escolha é que nunca fica claro o papel dos dados, já que a avaliação será dada entre gramáticas, assumindo que elas sejam capazes de gerar os dados da língua.

Essa escolha, por outro lado, permitiu a concepção de gramática como um sistema formal, o qual - dentro da Teoria Gerativa - foi sofrendo diversas reformulações e reinterpretações ao longo dos anos. Aqui vamos lidar com os fundamentos desse sistema formal, para que se possa entender o que exatamente é uma gramática gerativa.

Símbolos, Cadeias e Linguagens

As noções de símbolos, cadeias e linguagens são as noções fundamentais para entender a teoria da gramática como um sistema formal. E, sendo um sistema formal, ela pode descrever qualquer língua, não havendo um compromisso específico com as línguas naturais. De fato, esse sistema é muito útil para lidar a interpretação de linguagens de programação.

Símbolos (ou palavras) são os átomos das *cadeias*, sequências de símbolos. Representamos os símbolos de maneira abstrata por letras romanas minúsculas, como a, b, c, d. Um *alfabeto* é um conjunto de símbolos e é recebe como nome letras maiúsculas gregas como A, B, Γ, Δ

Em uma língua natural, cabe ao linguista estabelecer quais são os símbolos. Por exemplo, o alfabeto Π do português, considerando apenas a frase *eu comprei duas pizzas* seria $\Pi = \{eu, comprei, duas, pizzas\}$.

Cadeias recebem como nomes letras gregas minúsculas, como $\alpha, \beta, \gamma, \delta$. Uma frase de uma língua, também chamada de *sentença*, é uma cadeia. Logo, podemos dizer que $\alpha =$ "eu comprei duas pizzas" e que os símbolos de α pertencem ao conjunto S.

A um conjunto de cadeias damos o nome de *linguagem*. Uma *linguagem* pode ser uma conjunto finito ou infinito. Línguas naturais seriam um tipo de *linguagem*, um conjunto infinito de cadeias. Como dito anteriormente, uma cadeia α que pertença a uma linguagem L é uma sentença de L.

Gramáticas (Formais)

Uma gramática é a definição de uma linguagem, considerando que linguagem é um **conjunto** de cadeias. Dessa forma, uma gramática se dá nos termos das combinações dos símbolos do alfabeto de uma linguagem.

Uma gramática é formalmente definida como a quadrupla $G = (V, \Sigma, P, S)$, em que:

- V é um conjunto de todos os símbolos da linguagem
- Σ é o conjunto de símbolos terminais (alfabeto), $\Sigma \in V$
- P é o conjunto das produções (regras de combinação de símbolos)
- S é a raíz da gramática, $S \in V$

O conjunto de símbolos $V - \Sigma$ (V sem o Σ) é também chamado de N, conjunto de símbolos não-terminais.

A gramática é um sistema de substituições, de forma que, sentenças são formadas de duas formas:

- S é por si uma sentença.
- Seja $\alpha\rho\beta$ uma sentença, em que α, β e γ são cadeias de símbolos em V (terminais ou não-terminais) e $\rho \to \gamma$ é uma produção da gramática. A aplicação dessa produção, substituindo ρ por γ é uma sentença.

Hierarquias de Chomsky

Chomsky define, de uma maneira hierárquica, os tipos de gramática que podem existir em seu sistema formal, de acordo com as restrições de como pode ser o conjunto P de produções. Da menos para a mais restrita temos:

- Gramáticas Lineares
- Gramáticas Sensíveis ao Contexto
- Gramáticas Livres de Contexto
- Gramáticas Regulares

Gramáticas Lineares

As gramáticas lineares possuem produções $\alpha \to \beta$ que atendem as seguintes condições:

- $\alpha \in N$
- $\beta \in \Sigma \cdot N$ (linear à direita)
- $\beta \in N \cdot \Sigma$ (linear à esquerda)

Vamos considerar uma gramática G com as seguintes produções:

- $S \to aS$
- $S \to bS$
- $S \to cM$
- $M \to cJ$
- $J \to d$

Trata-se de uma gramática linear à direita, que gera sentenças de uma linguagem regular. Quais sentenças são geradas por G?

Gramáticas Livres de Contexto

Gramáticas livres de contexto possuem produções $\alpha \to \beta$ em que:

- $\alpha \in N$
- β é qualquer cadeia de símbolos terminais ou não terminais.

A gramática é livre de contexto pois o não terminal α pode ser substituido por β independentemente de seu contexto.

Gramáticas livres de contexto são utilizadas frequentemente para analisar linguagem natural. Por exemplo:

- $S \rightarrow SN \cdot SV$
- $SV \rightarrow V \cdot SN$
- $SV \rightarrow V$
- $SN \rightarrow Det \cdot N$
- $SN \rightarrow N$
- $N \rightarrow gato$
- $N \rightarrow queijo$
- $Det \rightarrow o$
- $V \rightarrow dormiu$
- $V \rightarrow comeu$
- $V \rightarrow beijou$

Gramáticas livres de contexto se diferenciam das lineares por apresentarem maior flexibilidade no componente β das produções. Quais sentenças podem ser geradas pelas produções da gramática acima?

Gramáticas Sensíveis ao Contexto:

Gramáticas sensíveis ao contexto possuem produções $\alpha A \beta \rightarrow \alpha \gamma \beta$ em que:

- α, β são quaisquer cadeias
- $A \in N$
- γ é qualquer cadeia

As gramáticas sensíveis ao contexto se diferenciam das livres de contexto por apresentarem um contexto α, β que regula a substituição de A por γ. Lembrando que tanto α como β podem ser vazios, o que faz as gramáticas livres de contexto serem um subconjunto das sensíveis ao contexto.

Praticamente todas as línguas naturais podem ser analisadas com uma gramática sensível ao contexto. O problema é que elas também possuem menos limitações do que uma língua natural. Acredita-se mais recentemente que as línguas naturais estão em algum lugar entre as gramáticas livres e sensíveis de contexto.

Gramáticas Irrestritas

Gramáticas irrestritas possuem produções $\alpha \rightarrow \beta$ em que:

- α, β são quaisquer cadeias

Gramáticas irrestritas podem gerar quaisquer tipos de linguagem.

Qual a importância dessa hierarquia?

Essas hierarquias são importantes quando se junta a elas o fato de que chomsky também previu qual tipo de recurso computacional seria necessário para criar gramáticas e reconhecedores de cada tipo de linguagem. Isso permitiu muito avanço na área de linguagens de programação.

É muito importante, no entanto, conhecer essas hierarquias pois elas estão na base dos formalismos gerativistas e muitos dos textos teóricos mais *pesados* irão se referir a elas.