Introdução ao Estudo Computacional da Linguagem

Marcelo Finger

Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

2º Semestre 2019

Tópicos

1. Preâmbulo
2. Linguagem e sua Modelagem
3. Sintaxe: A Análise da Estrutura
4. Linguagens Formais e Gramáticas Formais
5. Semântica
6. Pragmática

Tópicos

1. Preâmbulo
2. Linguagem e sua Modelagem
3. Sintaxe: A Análise da Estrutura
4. Linguagens Formais e Gramáticas Formais
5. Semântica
6. Pragmática

Estudo da Linguagem

- Ao menos, desde a antiguidade clássica
- Diversos enfoques
- Muitos deles eram dependentes de linguagens
- Visão unificadora a partir do século 19

Estudo da Linguagem

- Ao menos, desde a antiguidade clássica
- Diversos enfoques
- Muitos deles eram dependentes de linguagens
- Visão unificadora a partir do século 19
- Processamento automático a partir do século 20

Elementos da linguagem

- Sons

Elementos da linguagem

- Sons
- Ritmos (Prosódia)

Elementos da linguagem

- Sons
- Ritmos (Prosódia)
- Fonemas

Elementos da linguagem

- Sons
- Ritmos (Prosódia)
- Fonemas
- Palavras

Elementos da linguagem

- Sons
- Ritmos (Prosódia)
- Fonemas
- Palavras
- Sintagmas (Frases)

Elementos da linguagem

- Sons
- Ritmos (Prosódia)
- Fonemas
- Palavras
- Sintagmas (Frases)
- Significados

Elementos da linguagem

- Sons
- Ritmos (Prosódia)
- Fonemas
- Palavras
- Sintagmas (Frases)
- Significados
- Usos

Elementos da linguagem

- Sons – Acústica

Elementos da linguagem

- Sons – Acústica
- Ritmos (Prosódia)– Fonologia

Elementos da linguagem

- Sons – Acústica
- Ritmos (Prosódia)– Fonologia
- Fonemas – Fonética

Elementos da linguagem

- Sons – Acústica
- Ritmos (Prosódia)– Fonologia
- Fonemas – Fonética
- Palavras – Morfologia e Filologia

Elementos da linguagem

- Sons – Acústica
- Ritmos (Prosódia)– Fonologia
- Fonemas – Fonética
- Palavras – Morfologia e Filologia
- Sintagmas (Frases) – Sintaxe

Elementos da linguagem

- Sons – Acústica
- Ritmos (Prosódia)– Fonologia
- Fonemas – Fonética
- Palavras – Morfologia e Filologia
- Sintagmas (Frases) – Sintaxe
- Significados – Semântica

Elementos da linguagem

- Sons – Acústica
- Ritmos (Prosódia)– Fonologia
- Fonemas – Fonética
- Palavras – Morfologia e Filologia
- Sintagmas (Frases) – Sintaxe
- Significados – Semântica
- Usos – Pragmática

Várias Visões

Por exemplo, na sintaxe:

- Visão da estrutura sintagmática (constituintes)

Várias Visões

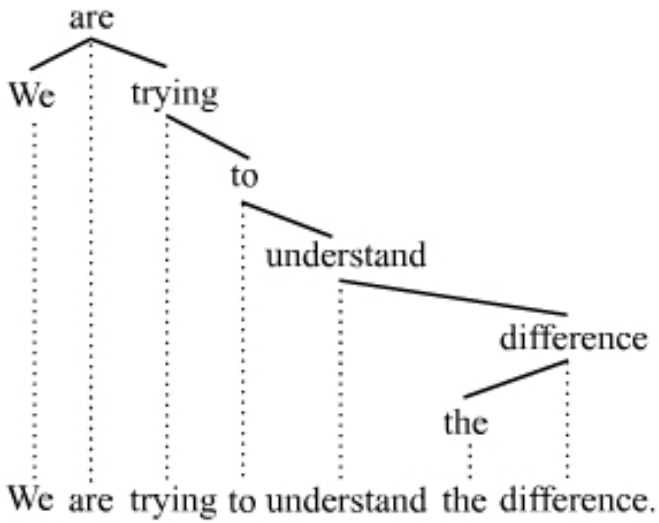
Por exemplo, na sintaxe:

- Visão da estrutura sintagmática (constituintes)
- Visão das dependências entre palavras [Séc XII]

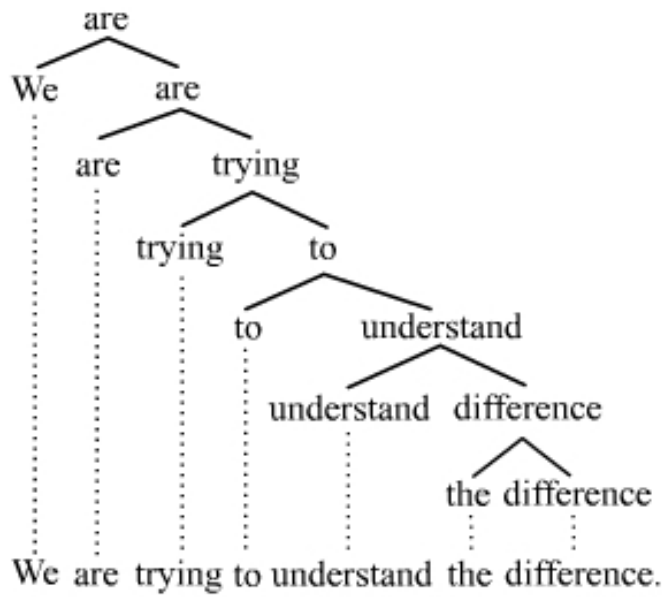
Várias Visões

Por exemplo, na sintaxe:

- Visão da estrutura sintagmática (constituintes)
- Visão das dependências entre palavras [Séc XII]

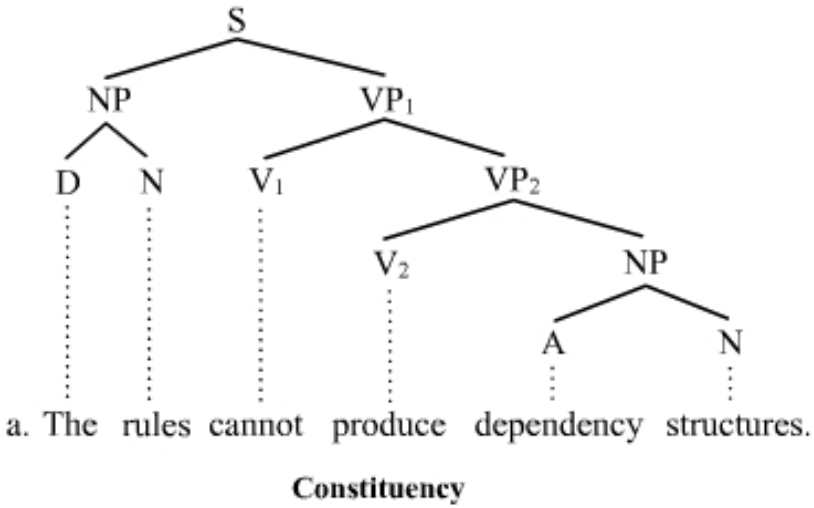


Dependency

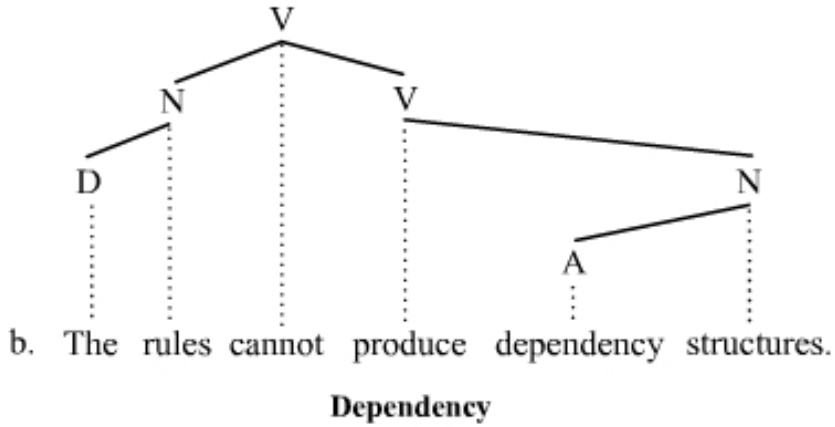


Constituency (BPS)

Estrutura: Sintagmática vs Dependência

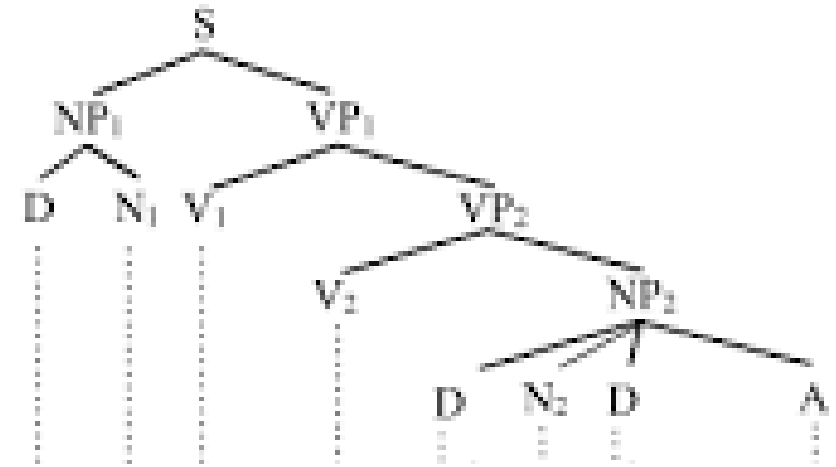


Constituency

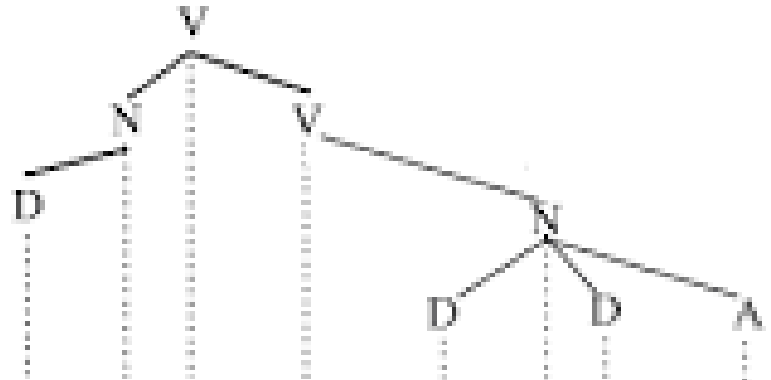


Dependency

Estrutura: Sintagmática vs Dependência



Esta árvore está ilustrando a relação de constituição



Esta árvore está ilustrando a relação de dependência

Tópicos

1 Preâmbulo

2 Linguagem e sua Modelagem

3 Sintaxe: A Análise da Estrutura

4 Linguagens Formais e Gramáticas Formais

5 Semântica

6 Pragmática

Linguagem

- Linguagem é um fenômeno humano.
- Desta forma, difere em natureza dos fenômenos *naturais*.
- Fenômenos naturais são estudados pela Física, Química e Biologia.
- Ciências Humanas, em geral, não possuem tratamento matemático.
- Ou melhor, não possuíam até bem recentemente.

Características da Lingüística

- É uma ciência eminentemente multidisciplinar:
  - Psicologia (cognição)
  - Ciência Física (emissão de som)
  - Biologia (anatomia humana influencia a fala)
  - Sociologia (Pragmática)
- Subdivide-se (grosseiramente) em:

  Fonética (sílabas), Fonologia (prosódia), Morfologia (palavras), Sintaxe (sentenças), Semântica (significado) e Pragmática (uso).

Modelagem Matemática da Lingüística

- Certas sub-áreas da lingüística podem ser modeladas matematicamente.
- Morfologia e Sintaxe: *Linguagens Formais e Autômatos*
- Sintaxe: *Lógica Categórica*
- Semântica: *Cálculo Lambda* e *Lógica Categórica*
- Estes modelos matemáticas não são numéricos ou quantitativos.
- Modelos numéricos e quantitativos são mais fáceis de processar
- Cf: Muita matemática precisou ser "inventada" para o desenvolvimento da Física.
- Idem deve ocorrer para a Linguística.

Tópicos

1 Preâmbulo

2 Linguagem e sua Modelagem

3 Sintaxe: A Análise da Estrutura

4 Linguagens Formais e Gramáticas Formais

5 Semântica

6 Pragmática

A Estrutura da Frase

- A *palavra* é unidade básica na análise da frase.
- A cada palavra é atribuída uma *categoria morfo-sintática*.
  Também conhecida como *classe gramatical*.
- Exemplos:

Categoria	*Conhecida como*	*Abreviação*
Determinante	Artigo	Det
Nome	Substantivo	N
Verbo Transitivo		VT

A estrutura da frase (cont.)

- Representação usual: $\langle Categoria \rangle \longrightarrow \langle Palavra \rangle$.
- Por exemplo, nas palavras da frase:

  o vilão beijou a mocinha

  Temos o seguinte *léxico*:

  $Det \longrightarrow$ o
  Det $\longrightarrow a$
  $N \longrightarrow$ vilão
  N $\longrightarrow mocinha$
  $VT \longrightarrow beijou$

Sintagmas

- Palavras se combinam para formar novos componentes.
- Estes componentes são chamados de **sintagmas**. Exemplos:

Sintagma	*Abreviação*
Sintagma Nominal	SN
Sintagma Verbal	SV
Sintagma Adjetival	SA
Sintagma Prepositivo	SP
Sentença	S

- Nota: sintagma em inglês é *phrase*.
- Sintagmas são *categorias sintáticas*.

Sintagmas (cont.)

- Genericamente, podemos dizer que categorias (sintáticas ou morfo-sintáticas) se combinam para gerar novas categorias.
- Desta forma, um conjunto de categorias

$$C_1, \ldots, C_n$$

  se combina para gerar uma categoria C.
- Representação: $C \longrightarrow C_1, \ldots, C_n$.
- Por exemplo:

  $S \longrightarrow SN\ SV$

  $SN \longrightarrow Det\ N$

  $SV \longrightarrow VT\ SN$

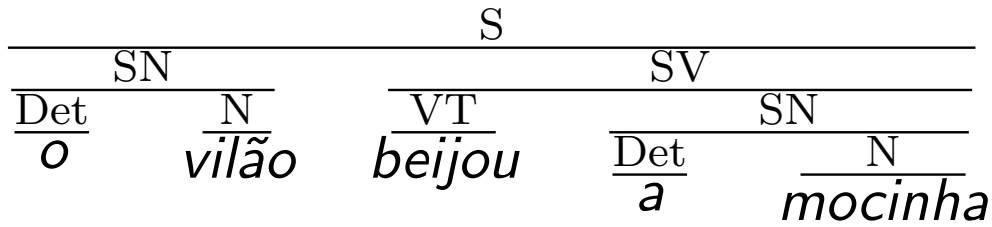
Sintagmas (cont.)

- Juntando todas as regras, temos uma **gramática** que gera (e reconhece) a frase:

o vilão beijou a mocinha

S $\longrightarrow$ SN SV	Det $\longrightarrow$ *a*
SN $\longrightarrow$ Det N	N $\longrightarrow$ *vilão*
SV $\longrightarrow$ VT SN	N $\longrightarrow$ *mocinha*
Det $\longrightarrow$ *o*	VT $\longrightarrow$ *beijou*

- Árvore sintática:



Tópicos

1 Preâmbulo

2 Linguagem e sua Modelagem

3 Sintaxe: A Análise da Estrutura

4 Linguagens Formais e Gramáticas Formais

5 Semântica

6 Pragmática

Linguagens Formais

- Seja V um conjunto não vazio de *símbolos* ou *palavras*.
- Uma *linguagem* $\mathcal{L}$ sobre V é um conjunto de cadeias formadas com os elementos de V:

$$\mathcal{L} \subseteq V^*$$

  onde, V^* é o conjunto de seqüências de 0 ou mais símbolos de V
- Se r e s são seqüências em V^*, rs representa a seqüência resultante da sua *concatenação*.
- ε é a seqüência vazia. Para todo s: $s\varepsilon = s = \varepsilon s$.

Exemplos de Linguagens Formais

- Exemplo. Com $V = \{a, b\}$ temos várias possíveis linguagens:
  - $L_1 = \{a, b, \varepsilon\}$
  - $L_2 = \{a^n, b^n | n = 0, 1, \ldots\}$
  - $L_3 = \{(ab)^n | n = 0, 1, \ldots\}$
- Se A e B são conjuntos, $AB = A \times B$.
- $A^2 = AA$, $A^3 = AAA$, etc.
- Por convenção: $A^0 = \{\varepsilon\}$.
- $\varnothing A = A\varnothing = \varnothing$
- Uma linguagem $\mathcal{L}_G \subseteq V^*$ poder ser definida por uma *gramática* G.

Gramáticas Formais

- Uma gramática G é uma quádrupla:

$$G = \langle V, T, P, S \rangle$$

onde:

- V é um conjunto não vazio, chamado *vocabulário*
- $T \subseteq V$ é um conjunto não vazio de *símbolos terminais*.
- $N = V - T$ é o conjunto de *símbolos não-terminais*.
- $S \in N$ é o símbolo inicial
- P é um conjunto finito de regras (*produções*) da forma

$$\alpha \longrightarrow \beta$$

onde $\alpha \in V^* N V^*$ e $\beta \in V^*$.

- Por exemplo, a gramática vista anteriormente é formado dos seguintes elementos:

$$T = \{o,\ vilão,\ beijou,\ a,\ mocinha\}$$
$$N = \{\mathrm{S},\ \mathrm{SN},\ \mathrm{SV},\ \mathrm{Det},\ \mathrm{N},\ \mathrm{VT}\}$$
$$V = T \cup N$$
$$P = \begin{cases} S \longrightarrow SN\ SV \\ SN \longrightarrow Det\ N \\ SV \longrightarrow VT\ SN \\ Det \longrightarrow o \\ Det \longrightarrow a \\ N \longrightarrow vilão \\ N \longrightarrow mocinha \\ VT \longrightarrow beijou \end{cases}$$

Classificação de Gramáticas Formais

- Gramáticas são classificadas tipo restrições que se aplica sobre as regras de produção admitidas na gramática:

$$\alpha \longrightarrow \beta$$

- Notação: $L(G)$ é a linguagem reconhecida/gerada pela gramática G.

Gramáticas regulares

Gramáticas regulares admitem apenas produções onde:

- α é um não terminal ($\alpha \in N$)
- $\beta \in T^*NT^* \cup T^*$, ou seja, β pode conter no máximo um não terminal, rodeado por 0 ou mais não-terminais.

Gramáticas regulares são muito úteis para o reconhecimento padrões em textos, utilizados em linguagens de processamento de texto como Perl.
Gramáticas regulares são reconhecíveis por *autômatos finitos*, que possuem implementação eficiente.

Expressões Regulares

Seja T um vocabulário (terminais). Definimos *expressões regularess* sobre T, $ER(T)$:

- $\epsilon \in ER(T)$;
- $v \in T \Rightarrow v \in ER(T)$;
- $x \in ER(T) \Rightarrow (x)^* \in ER(T)$;
- $x, y \in ER(T) \Rightarrow xy \in ER(T), \quad x|y \in ER(T)$.

Define-se ainda

- $x^+ = x(x)^*$;
- $[v_1 \ldots v_n] = (v_1|\ldots|v_n) \;\; (v_i \in T)$;
- $[\,\hat{}\,v_1 \ldots v_n] = (v_{i_1}|\ldots|v_{i_m}) \;\; (v_{i_j} \in T - \{v_1, \ldots, v_n\})$.

Expressões Regulares (exemplos)

- $T = \{a, b, c\}$.
- $L(E) =$ Linguagem reconhecida pela e.r. $E \in ER(T)$.
- $L(a|b|c) = T$.
- $L((ab)^*) = \{\epsilon, ab, abab, ababab, \ldots\}$.
- $L((a|b)^*) = \{\epsilon, a, b, aa, ab, ba, bb, \ldots\}$
- $L((a|b)^*c) = xc, x \in L((a|b)^*) =$
  $\{c, ac, bc, aac, abc, bac, bbc, \ldots\}$
- Note que $L((ab)^*) = L((ab(ab)^*)^*)$, ou seja, mais de uma e.r. representa a mesma linguagem.

Expressões Regulares e Gramáticas Regulares

Theorem

Existe uma e.r. E se e somente se existe uma gramática regular G_r tal que $L(E) = L(G_r)$.

Exemplos:

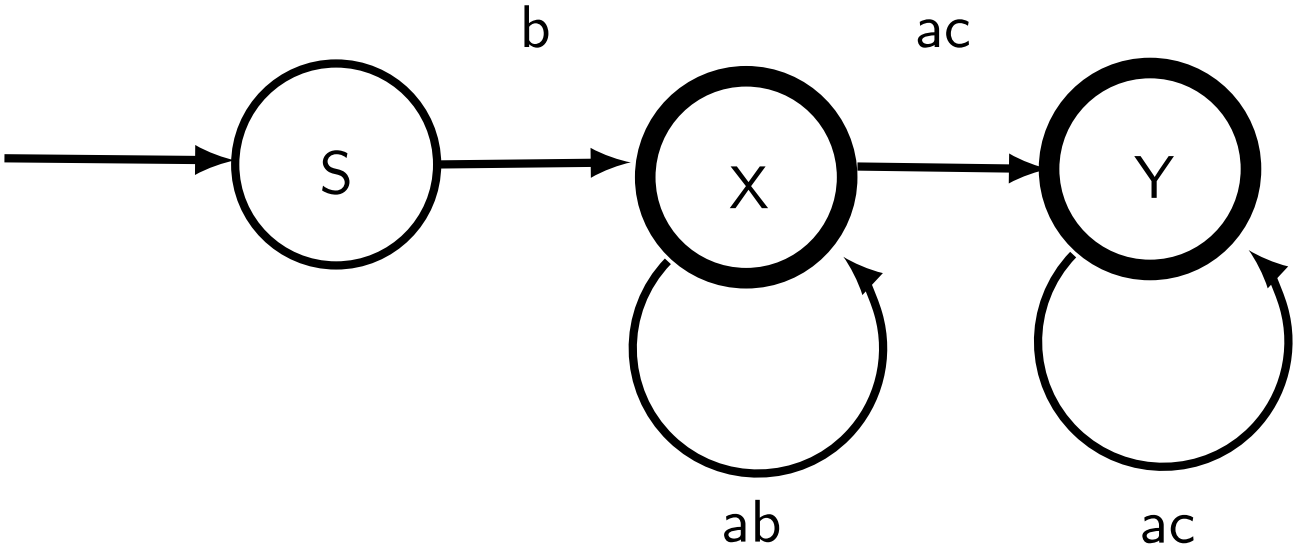
$$(a|b)^*c \quad \Leftrightarrow \quad \begin{cases} S \to C \\ S \to aS \\ S \to bS \\ C \to c \end{cases}$$

$$b(ab)^*(ac)^* \quad \Leftrightarrow \quad \begin{cases} S \to bX \\ X \to Y \\ X \to abX \\ Y \to \epsilon \\ Y \to acY \end{cases}$$

Autômatos Finitos

- Um autômato finito é uma quíntupla $A = (Q, T, M, q_0, F)$ onde:
  - $Q \neq \varnothing$ é um conjunto de estados;
  - T é um vocabulário (terminais);
  - $M : Q \times T \rightarrow Q$ é a função de *transição de estados*;
  - $q_0 \in Q$ é o estado inicial
  - $F \subseteq Q$ são os estados de aceitação
- Uma cadeia $s \in T^*$ é aceita por A se, a partir do estado inicial, seguindo a função de transição, termina-se num estado de aceitação.
- $L(A)$ é a linguagem reconhecida/gerada por A.

Autômatos Finitos (Exemplo)



Note que $L(A) = L(b(ab)^*(ac)^*)$.

Autômatos Finitos e Gramáticas Regulares

- Existe um autômato A sse existe uma gramática regular G tal que $L(A) = L(G)$.
  Logo, há equivalência entre a.f.'s , e.r.'s e gramáticas regulares.
  Vide exemplo anterior.
- O autômato é *determinístico* se não possui arcos rotulados por ϵ (ou se M é uma função, e não uma relação).
- Para todo a.f. não-determinístico A existe um a.f. determinístico A' com $L(A) = L(A')$.
- Um a.f. pode ser minimizado, gerando-se um a.f. equivalente com menos estados.
- O a.f. mínimo é único (a menos de isomorfismos).

Autômatos Acíclicos

- Autômatos *acíclicos* reconhecem linguagens *finitas*.
- Um *léxico* é uma linguagem finita. Pode ser minimizado em um autômato finito.
- Forma muiiiiito eficiente de se guardar/indexar grandes quantidades de dados.
- Léxicos inteiros podem ser armazenados em memória.

Gramáticas Livres de Contexto

Gramáticas livres de contexto (GLC) admitem apenas produções $\alpha \longrightarrow \beta$ onde:

- α é um não-terminal ($\alpha \in N$)
- $\beta \in V^*$, ou seja, β pode conter uma seqüência de elementos do vocabulário.

As GLCs são de grande utilidade prática.
Basicamente, TODAS as linguagens de programação são definidas por GLCs. Os *tipos* das variáveis são a única condição de contexto utilizadas em linguagens de programação, em geral.
GLCs são reconhecidas por *autômatos de pilhas*, também de eficiente implementação no computador.

A seguir, apresentamos um trecho de uma gramática GLC de uma linguagem pseudo-Pascal, com símbolo inicial Comando:

Comando $\longrightarrow$ Var := Expressão

Comando $\longrightarrow$ If Cond Comando Else Comando

Comando $\longrightarrow$ While Cond Comando

Comando $\longrightarrow$ Bloco

Bloco $\longrightarrow$ Begin SeqComandos End

SeqComandos $\longrightarrow$ Comando

SeqComandos $\longrightarrow$ Comando ; SeqComandos

$\vdots$

Gramáticas Sensíveis ao Contexto

Gramáticas Sensíveis ao Contexto admitem apenas produções da forma:

$$A\ X\ B \longrightarrow A\ Y\ B \text{ ou}$$
$$S \longrightarrow \varepsilon$$

onde $A, B \in V^*$, $X \in N$ e $Y \in V^+$.
A sensibilidade ao contexto aumenta muito a complexidade de se reconhecer tais gramáticas.

Hierarquia de Chomsky

Gramáticas Genéricas	: tipo 0
Gramáticas Sensíveis a Contexto	: tipo 1
Gramáticas Livres de Contexto	: tipo 2
Gramáticas Regulares	: tipo 3

Referência: *Introduction to Formal Languages*, G. E. Révész, 1983. Dover Publications (1991)

Geração *vs* Reconhecimento

Conceitos duais:

Geração: Dada uma gramática G e uma sentença s, s é gerada por esta gramática?

Reconhecimento: Um programa que responde corretamente a pergunta acima é um reconhecedor para aquela gramática.

A saída de um reconhecedor é:

- NÃO, se G não gera s.
- SIM, se G gera s. Neste caso, podemos ter como saída uma *árvore sintática* associada a s.

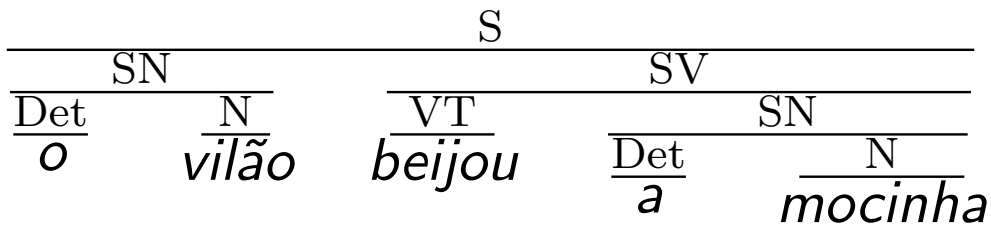
Pode haver mais de uma árvore sintática associada a um s.

Árvores de Reconhecimento

A sentença: *o vilão beijou a mocinha*
ao ser reconhecida pela gramática:

S $\longrightarrow$ SN SV	SN $\longrightarrow$ Det N
SV $\longrightarrow$ VT SN	Det $\longrightarrow$ *o*
Det $\longrightarrow$ *a*	N $\longrightarrow$ *vilão*
N $\longrightarrow$ *mocinha*	VT $\longrightarrow$ beijou

gera a seguinte árvore sintática:



Gramáticas Ambíguas

Uma mesma sentença gera mais de uma árvore sintática.

Exemplo: *Eu vi o menino com o telescópio.*

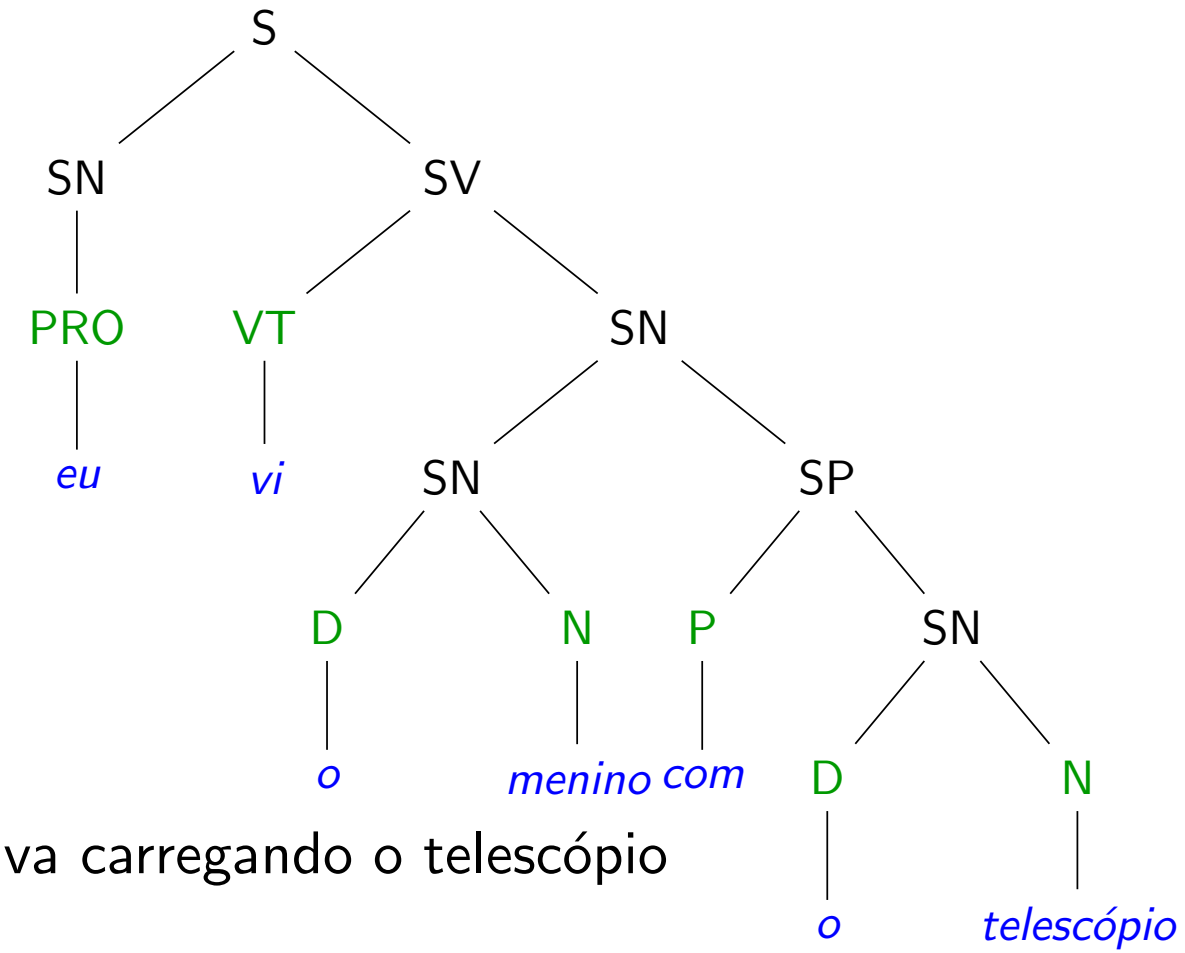
Gramática:

$$\begin{aligned}
S &\longrightarrow SN\ SV \\
SN &\longrightarrow Det\ N \\
SN &\longrightarrow PRO \\
SN &\longrightarrow SN\ SP \\
SV &\longrightarrow VT\ SN \\
SV &\longrightarrow SV\ SP \\
SP &\longrightarrow P\ SN
\end{aligned}$$

$$\begin{aligned}
PRO &\longrightarrow \textit{eu} \\
VT &\longrightarrow \textit{vi} \\
Det &\longrightarrow \textit{o} \\
P &\longrightarrow \textit{com} \\
N &\longrightarrow \textit{telescópio} \\
N &\longrightarrow \textit{menino}
\end{aligned}$$

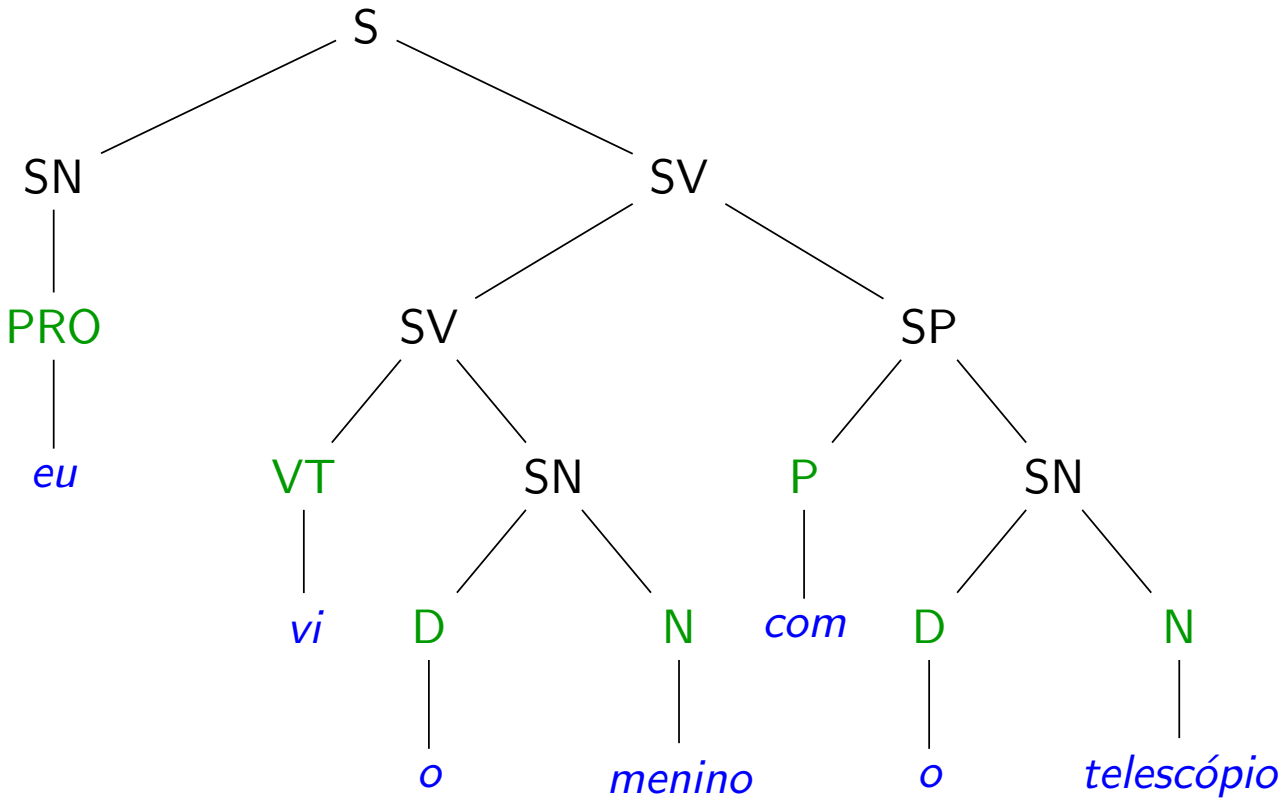
Nota: SP é o sintagma prepositivo (adjunto)

Sentido 1: Anexação do SP ao SN



Sentido: o menino estava carregando o telescópio

Sentido 2: Anexação do SP ao SV



Sentido: eu usei o telescópio para ver o menino

Tópicos

1 Preâmbulo

2 Linguagem e sua Modelagem

3 Sintaxe: A Análise da Estrutura

4 Linguagens Formais e Gramáticas Formais

5 Semântica

6 Pragmática

Interface Sintaxe-Semântica

Fenômeno: a cada árvore sintática atribui-se um significado (potencialmente) distinto.
Problema:

- Como atribuir significado a uma árvore sintática?
- O que é "significado"?

Semânticas Formais

- Representação matemática formal do significado
- **Composicionalidade:** O significado de um elemento é uma composição dos significados de suas partes
- Gramáticas Categóricas baseadas em Lógicas Categóricas e Cálculo Lambda
- Ponte formal para a interface Sintaxe-Semântica

Semânticas Contextuais

- Expressões com significados semelhantes ocorrem m contextos semelhantes
- O contexto é a semântica, a semântica é um padrão
- Inserção (embedding) em espaços vetoriais n-dimensionais
- Pontos próximos representam conceitos semanticamente semelhantes
- Exploram propriedades dos espaços vetoriais no processamento

Tópicos

1 Preâmbulo

2 Linguagem e sua Modelagem

3 Sintaxe: A Análise da Estrutura

4 Linguagens Formais e Gramáticas Formais

5 Semântica

6 Pragmática

Pragmática

Relação entre a linguagem e seus falantes.

Pragmática trata do uso concreto da linguagem, enquanto a semântica e a sintaxe constituem a construção teórica

Trata de assuntos como deixis, revezamento na conversação, organização do texto, pressuposições e implicatura.

Interface ente linguagem e cultura

Exemplos de Pragmática em Ação

1. Eu gostaria de saber se você teria condições de me dar uma carona.
2. Que beleza está o trânsito hoje!
3. Está um calor aqui dentro, não?
4. Que bela loja você tem, seria uma pena se algo acontecesse a ela.