**C**ollaborative

**A**cademic

**T**ext

**A**dvisor

# Um verificador de **estilo** para textos acadêmicos de **Computação**

*Ana Luiza Domingues Fernandez Basalo*
Orientação *Professor Marco Aurélio Gerosa*

**alunos** de Computação
e a **escrita** de
**textos acadêmicos**

problemas na escrita

- levam orientadores a **empregar muito tempo** com correção de textos
- **prejudicam** a **disseminação** dos resultados das pesquisas

## Erros **ortográficos**

A **caza** era ampla e arejada.

## Erros **gramaticais**

Eles **vai** ao estádio.

Problemas de **estilo**

(em **textos acadêmicos** de **Computação**)

• • •

“testes unitários”

“o algoritmo **retorna** um determinado valor”

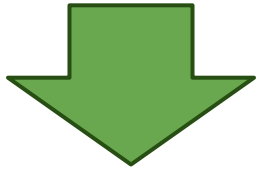
“**rodar** um programa”

Dado um
texto acadêmico de Computação
como entrada,

antes de
"saírmos à **cata**"
de problemas e vícios de estilo...

# Segmentação (*Tokenization*)

As meninas são inteligentes.



| As | meninas | são | inteligentes | . |
|---|---|---|---|---|

debugamos

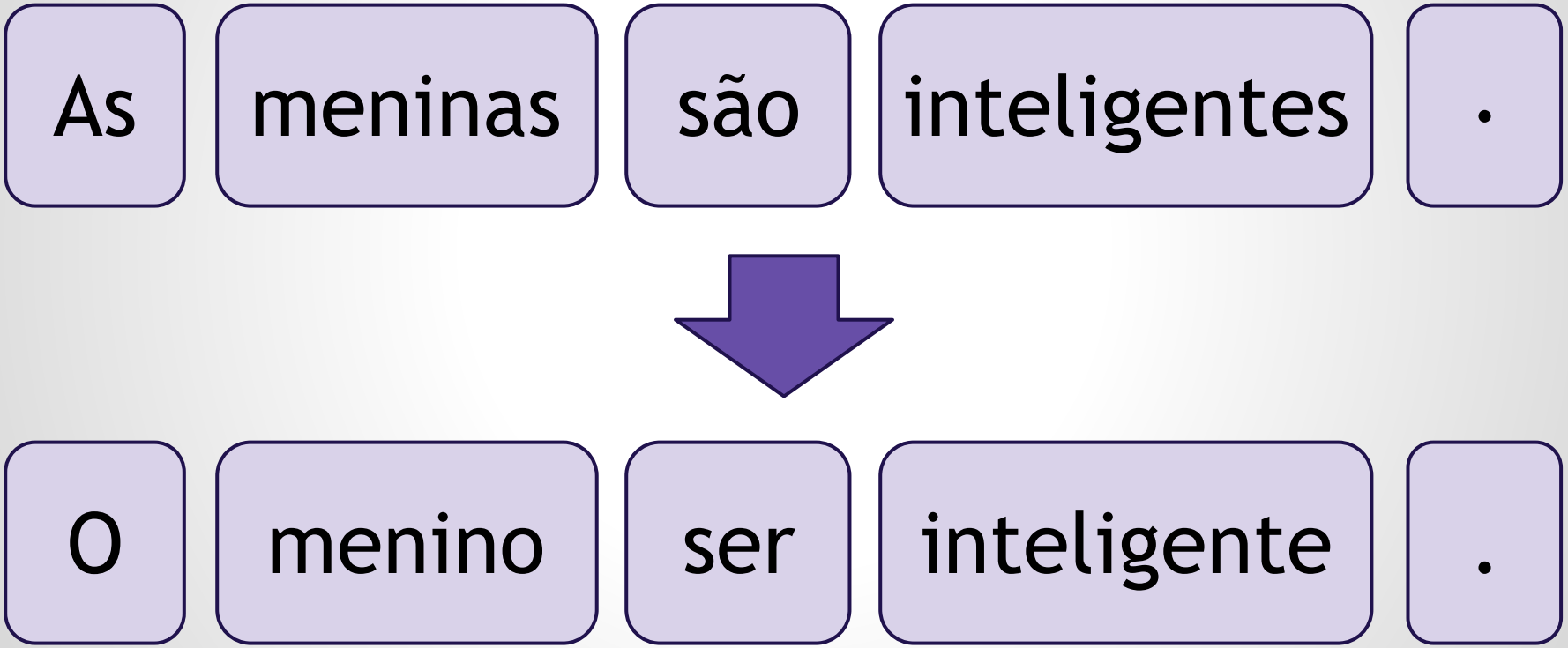
debuguei

# debugar

debugarei

debugam

# Lematização (*Lemmatization*)



como

Sou vegetariano, não **como** carne.

O aluno fez **como** o professor mandou.

- **POS-Tagging**
  (*Part-of-speech tagging*)

- **Modelo estatístico** para etiquetar palavras e obter lemas

Agora **sim** podemos realizar a **busca** de problemas de estilo

**Encontrar** problemas de estilo

=

**busca de padrões**

(para este trabalho)

- Formalmente: busca de um **conjunto finito** de padrões
- Algoritmo **Aho-Corasick**

# Aho-Corasick

Dado um conjunto finito de padrões (um *"dicionário"*), constrói uma máquina de estados. Dado um texto, a busca pelos padrões será realizada usando este **autômato finito**.

# Exemplo

*dicionário*

{ a, ab, bc, bca, c, caa }

# Função *"goto"*

Começamos com um **grafo dirigido** que possui **um único vértice** - o **estado inicial.**

# Função *"goto"*

{ **a**, ab, bc, bca, c, caa }

Começamos com "a".

# Função *"goto"*

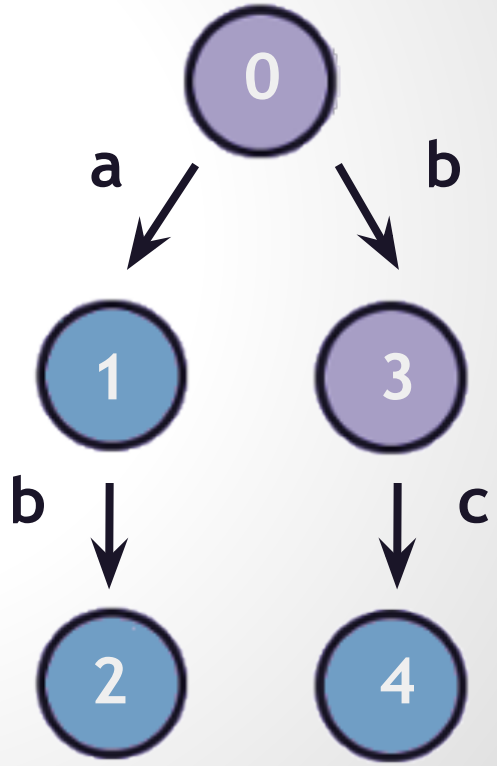
{ a, **ab**, bc, bca, c, caa }

Inserção de "ab".

# Função *"goto"*

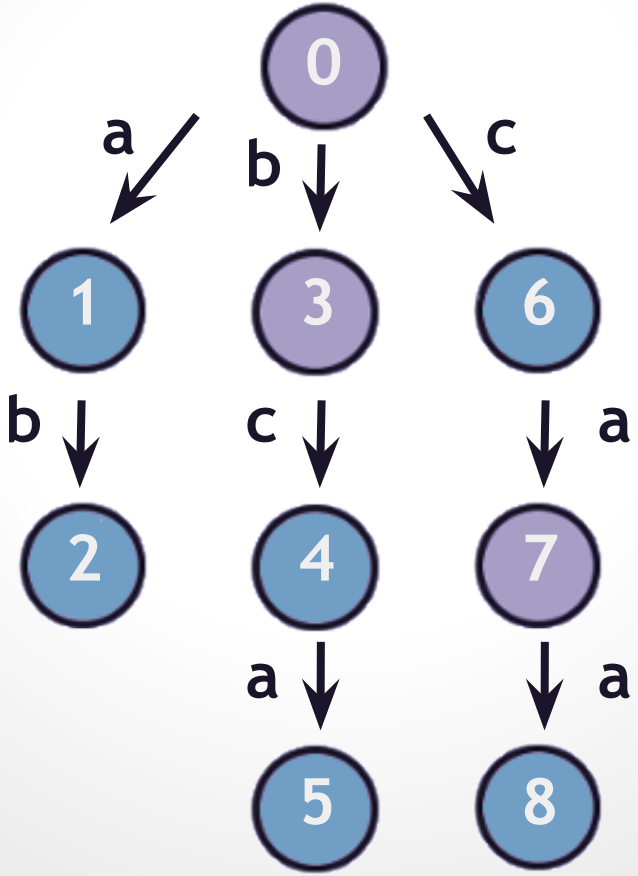
{ a, ab, **bc**, bca, c, caa }

Inserção de "bc".



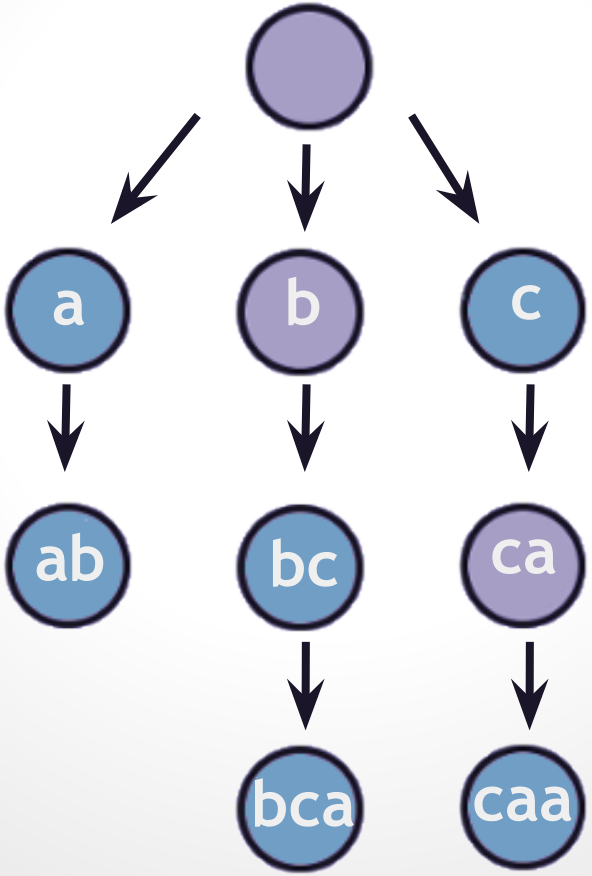
# Função *"goto"*

{ a, ab, bc, bca, c, caa }



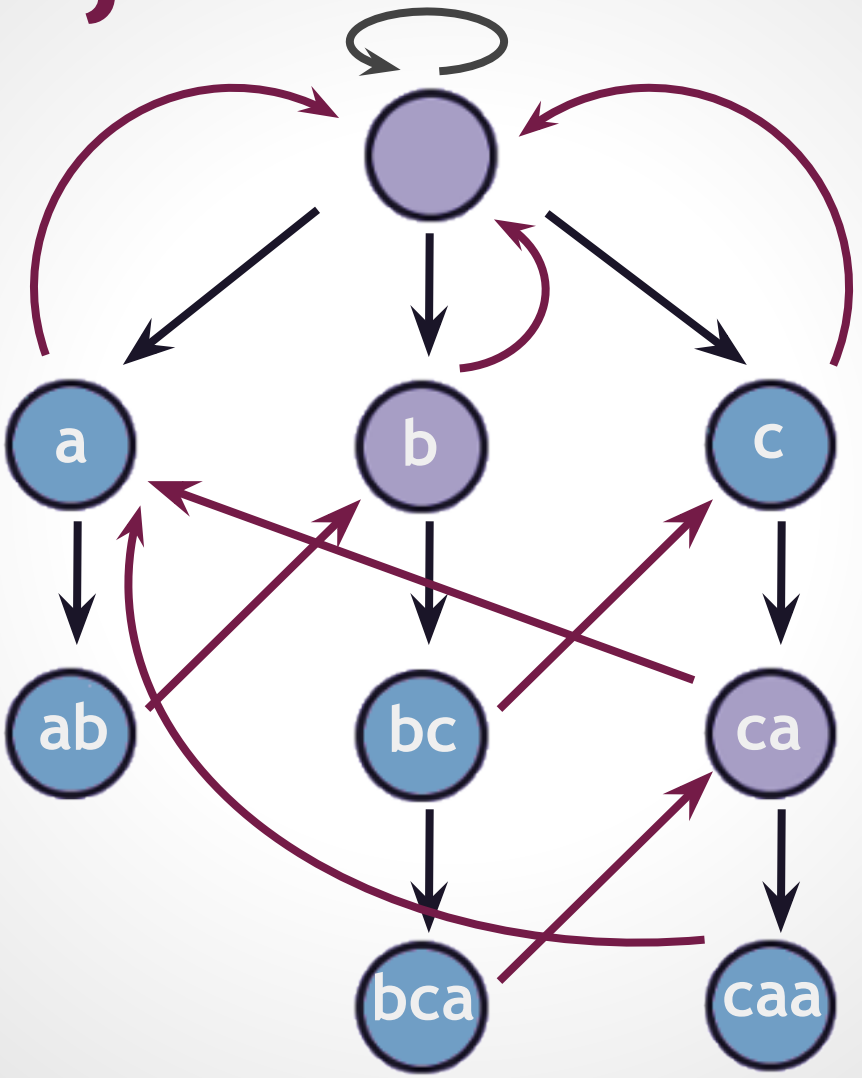
# Função *"output"*

{ a, ab, bc, bca, c, caa }



Temos uma **árvore**. Mas queremos um **autômato finito determinístico**.

# Função *"fail"*

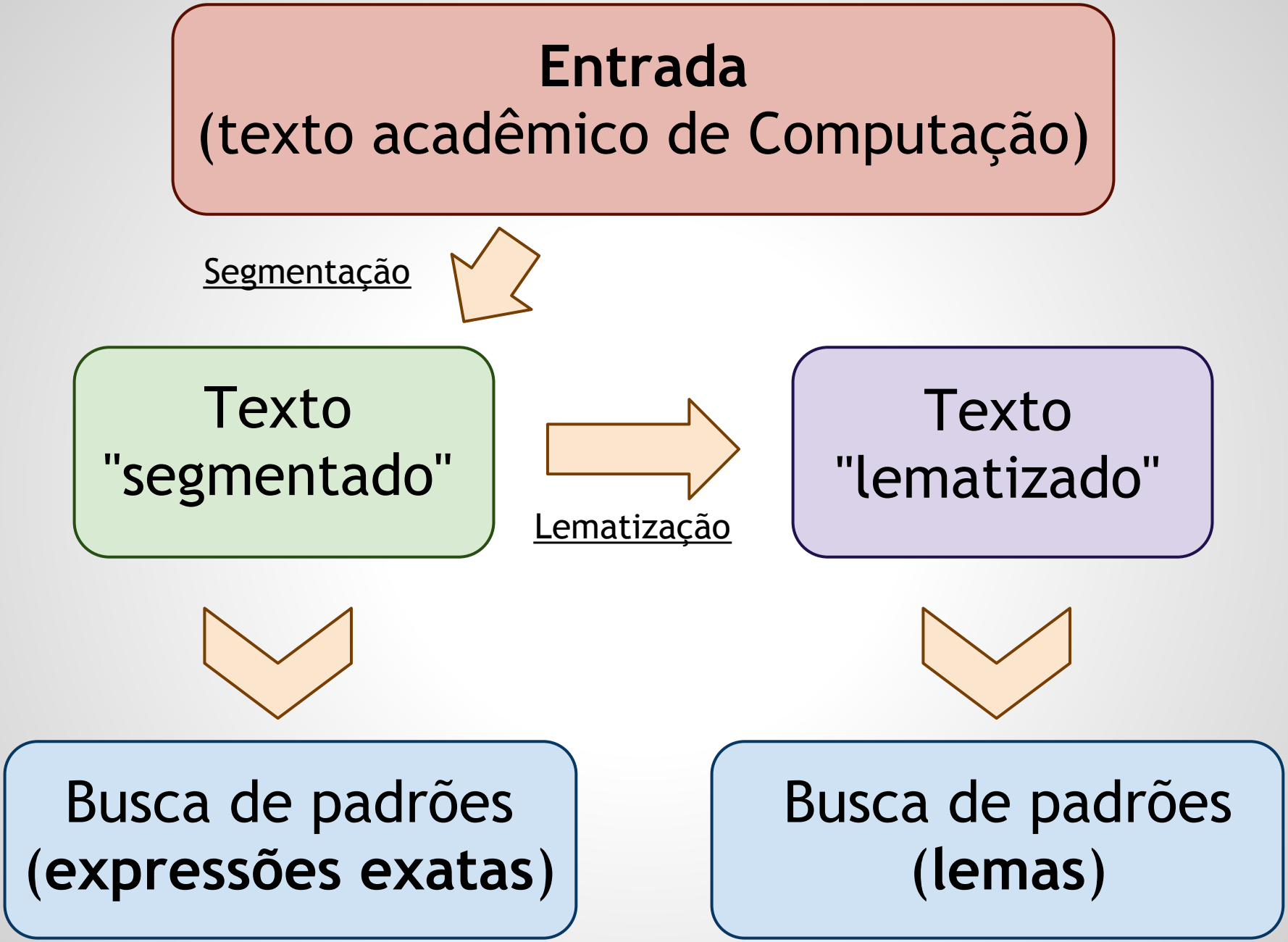


# Busca dos padrões

- "Consome" os caracteres e "passeia" no grafo dirigido
- Note que o algoritmo "percorre" o texto de entrada **uma única vez**

# Recapitulando

...


**Entrada**
(texto acadêmico de Computação)
Segmentação
Texto "segmentado"
Lematização
Texto "lematizado"
Busca de padrões (**expressões exatas**)
Busca de padrões (**lemas**)

Está pronto.

**Não**, não está.

cores

*cores* = núcleos

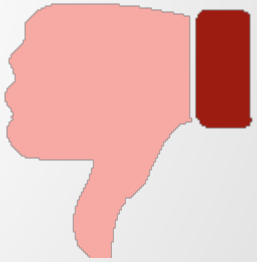
(problema de estilo)

cores = plural de "cor"

(tudo certo aqui)

# rodar

- Aspectos semânticos
- Inteligência Coletiva
- Sistema Colaborativo



- Extração de **palavras-chave**
- Comparações usando uma **escala de similaridade**
- **Decide** se exibe ou não a sugestão de melhoria de estilo

Agora sim, está pronto.

# Demonstração

# Conclusão

- A escrita é a principal forma de **comunicação científica**, devemos ter cuidado ao elaborar textos acadêmicos.

- É viável o desenvolvimento de sistemas computacionais para **automatizar**, ainda que em parte, a **verificação de estilo** de textos.

# Obrigada. =]

**MAC0499** - Trabalho de Formatura Supervisionado

**Instituto de Matemática e Estatística**
Universidade de São Paulo

16 de novembro de 2011