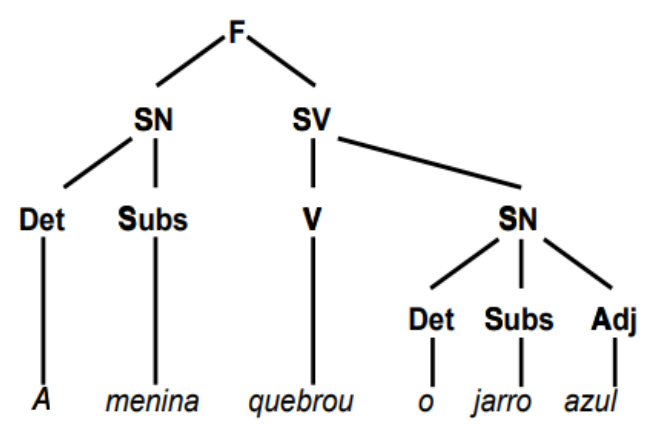
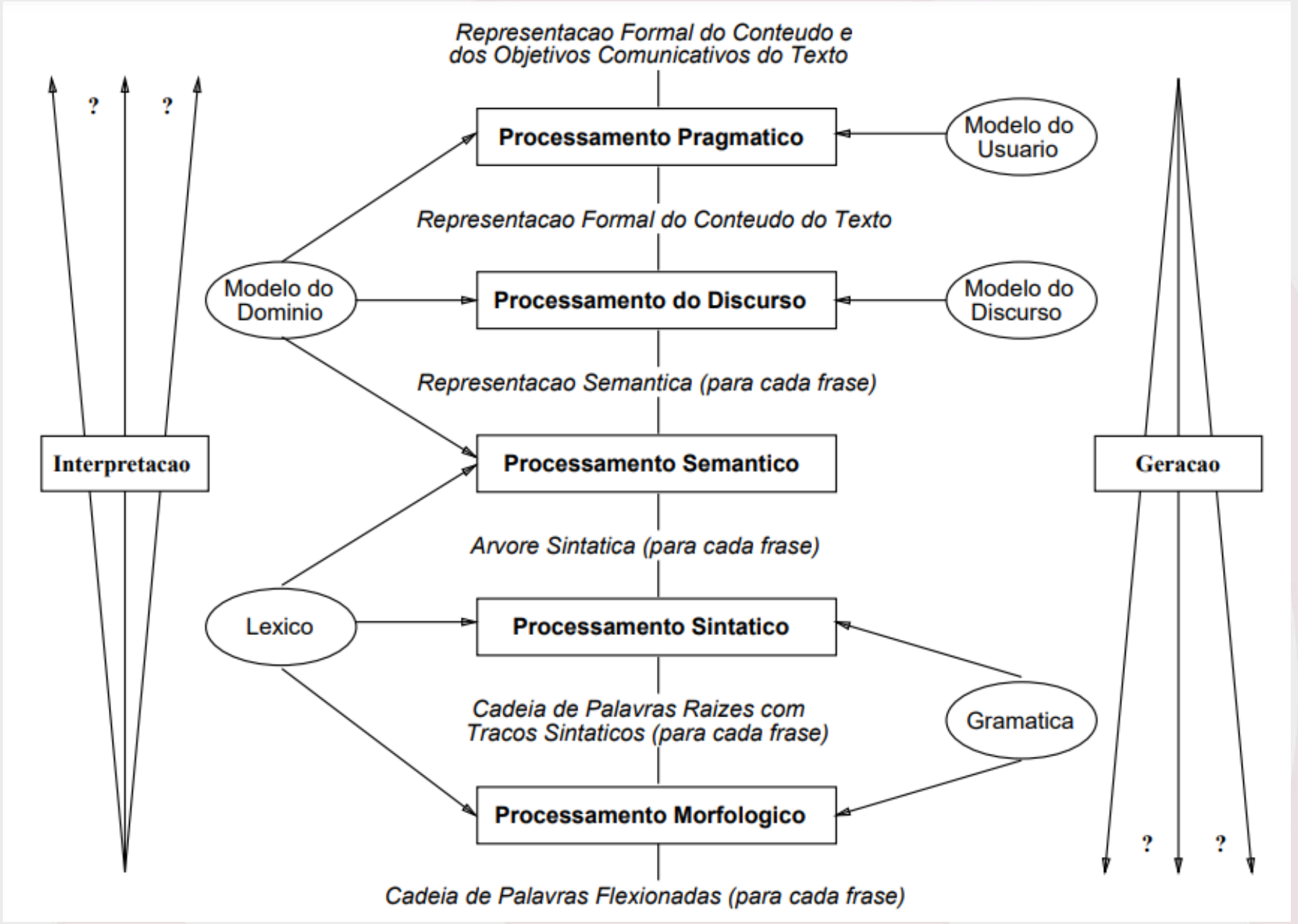
# Introdução

Árvore sintática



# Introdução

- **Corpus**:
  - Um grande conjunto de documentos em linguagem natural.
  - Base de dados para acumular estatísticas textuais e facilitar a análise de um determinado tipo de texto.
- **Gramática**:
  - Estratégica para tratar formalmente uma linguagem natural.
  - É uma especificação matemática da estrutura das sentenças de uma linguagem.

# Introdução

- **Gramática**:
  - S = Símbolo inicial.
  - T = Símbolos terminais.
  - N = Símbolos não terminais.
  - R = Regras de Produção.

$S$ = {frase}
$\mathcal{T}$ = {o, gato, rato, caçou}
$\mathcal{N}$ = {frase, sujeito, predicado, artigo, substantivo, verbo}
$\mathcal{R}$ = {frase --> sujeito, predicado ;
sujeito --> artigo, substantivo;
predicado --> verbo, artigo, substantivo;
artigo --> [o];
substantivo --> [gato] | [rato];
verbo --> [caçou]}

# Introdução

- **Léxico**:
  - Um léxico é uma coleção de informações sobre as palavras de uma linguagem sobre as categorias lexicais às quais elas pertencem.
  - Pode incluir características:
    - Morfológicas (conjugação de verbos, inflexão de substantivos etc.).
    - Sintáticas (categoria gramatical, regência verbal etc.).
    - Semântica (conceitos do domínio da aplicação).
  - Exemplo:
    - OpLexicon:
      - Léxico de sentimento com cerca de 15.000 palavras polarizadas.
      - Classificadas por sua categoria morfológica.
      - Polaridades positivas, negativas e neutras.

# Introdução

- **Léxico**:
  - Exemplo:

```
mesa
    <categoria> = substantivo
    <gênero> = feminio
    <número> = singular

comprou
    <cat> = verbo
    <tempo> = pretérito-perfeito
    <número> = singular
    <pessoa> = 3
    <arg1> = SN
    <arg2> = SN
```

# Questões de Concurso

- Prova: CESPE / CEBRASPE - 2021 - SEFAZ-CE - Auditor Fiscal de Tecnologia da Informação da Receita Estadual
- Um dos desafios do processamento de linguagem natural (PLN) é a polissemia, ou seja, a característica de palavras e frases poderem ter mais de um significado.

# Questões de Concurso

- Prova: CESPE / CEBRASPE - 2021 - SEFAZ-CE - Auditor Fiscal de Tecnologia da Informação da Receita Estadual
- Um dos desafios do processamento de linguagem natural (PLN) é a polissemia, ou seja, a característica de palavras e frases poderem ter mais de um significado.

# Questões de Concurso

- Prova: FCC - 2019 - TRF - 4ª REGIÃO - Analista Judiciário - Sistemas de Tecnologia da Informação
- Um Analista necessita desenvolver uma aplicação chatbot que simula um ser humano na conversação com as pessoas. Para isso o Analista deve usar pesquisa em Processamento de Linguagem Natural – PLN que envolve três aspectos da comunicação, quais sejam,
- A Som, ligado à fonologia, Estrutura que consiste em análises morfológica e sintática e Significado que consiste em análises semântica e pragmática.
- B Áudio, ligado à fonologia, Estrutura que consiste em análises de línguas estrangeiras e Significado que consiste em análises semântica e pragmática.
- C Conversação, ligado à tecnologia de chatbot, Semântica que consiste em análises de línguas estrangeiras e Arquitetura Spelling que realiza as análises sintática e pragmática.
- D Business Intelligence, ligado à tecnologia OLAP, Mining que consiste em análises de línguas em geral e Spelling que realiza as funções de chatbot.
- E Áudio, ligado à fonologia, Estrutura que consiste em análises semântica e pragmática e Significado que consiste em análise das línguas em geral.

# Questões de Concurso

- Prova: FCC - 2019 - TRF - 4ª REGIÃO - Analista Judiciário - Sistemas de Tecnologia da Informação
- Um Analista necessita desenvolver uma aplicação chatbot que simula um ser humano na conversação com as pessoas. Para isso o Analista deve usar pesquisa em Processamento de Linguagem Natural – PLN que envolve três aspectos da comunicação, quais sejam,
- A Som, ligado à fonologia, Estrutura que consiste em análises morfológica e sintática e Significado que consiste em análises semântica e pragmática.
- B Áudio, ligado à fonologia, Estrutura que consiste em análises de línguas estrangeiras e Significado que consiste em análises semântica e pragmática.
- C Conversação, ligado à tecnologia de chatbot, Semântica que consiste em análises de línguas estrangeiras e Arquitetura Spelling que realiza as análises sintática e pragmática.
- D Business Intelligence, ligado à tecnologia OLAP, Mining que consiste em análises de línguas em geral e Spelling que realiza as funções de chatbot.
- E Áudio, ligado à fonologia, Estrutura que consiste em análises semântica e pragmática e Significado que consiste em análise das línguas em geral.

# Captura de Dados

- **Rastreamento de dados (*crawling*)**:
  - Utilizada por mecanismos de buscas e extratores de dados.
  - Fases:
    - *Crawler* obtém lista de URL a partir de uma URL base.
    - Extração dos dados existentes nessas URL.
- **Raspagem de dados (*scraping*)**:
  - Extração automática eficiente dos conjuntos de dados estruturados (principalmente HTML).
  - O *Web Scraper* acessa páginas web, encontra elementos de dados especificados na página e os extrai.
  - O *Web Scraper* imita a operação de um usuário em busca de informações em um navegador, automatizando e acelerando este acesso.

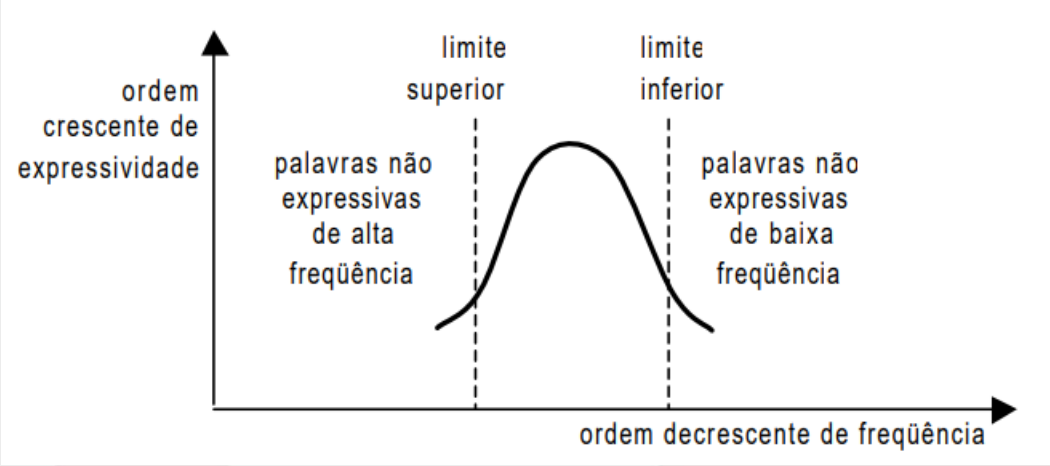
# Pré-processamento

- **Normalização**:
  - Tokenização, transformação de letras maiúsculas para minúsculas, remoção de caracteres especiais, remoção de tags HTML/Javascript/CSS.
  - O processo de tokenização tem como objetivo separar palavras ou sentenças em unidades.

| Notícia | Notícia Processada em Tokens |
|---|---|
| "Gustavo Pedreira Ferraz,<br>que admitiu buscar<br>malas de dinheiro<br>para Geddel Vieira Lima,<br>afirmou à PF ter trabalhado<br>na campanha Presidencial de 2014<br>para o então candidato Aécio Neves." | "Gustavo", "Pedreira", "Ferraz",<br>"que", "admitiu", "buscar"<br>, "malas", "de","dinheiro",<br>"para", "Geddel","Vieira ", "Lima ",<br>",", "afirmou","à","PF","ter","trabalhado",<br>"na","campanha","Presidencial","de", "2014"<br>,"para", "o", "então", "candidato", "Aécio", "Neves" |

# Pré-processamento

- **Remoção de *stopwords***:
  - Consiste em remover palavras muito frequentes, tais como “a”, “de”, “o”, “da”, “que”, “e” e “do”.
  - São removidas palavras não relevantes para o problema de PLN.
- **Análises estatísticas:**



# Pré-processamento

- **Correção Ortográfica:**
  - Corrige erros de digitação, abreviações e vocabulário informal.
- **Etiquetagem (*part-of-speech tagger*)**:
  - Colocação de etiqueta em cada palavra para identificar a categoria gramatical ou elementos morfológicos ou semânticos.

# Pré-processamento

- **Remoção de numerais**:
  - Remove-se os números e palavras/símbolos como “R$”, “$”, “US$”, “kg”, “km”, “milhões”, “bilhões” etc.
- **Stemização ou Lematização:**
  - Reduzir uma palavra ao seu radical.
  - As palavras “gato”, “gata”, “gatos” e “gatas” reduziriam-se para “gat”.
  - As palavras “tiver”, “tenho”, “tinha”, “tem” são formas do mesmo lema “ter”.

| *Word* | *Stem* | *Word* | *Stem* |
|---|---|---|---|
| magically | magic | groveling | grovel |
| chewing | chew | painful | pain |
| unequal | unequ | daguerreotype | daguerreotyp |
| shoddiness | shoddi | magnitude | magnitud |
| headline | headlin | standing | stand |
| ruinously | ruinous | obstruction | obstruct |
| allergenic | allergen | bagpiper | bagpip |
| signified | signifi | disunite | disunit |
| truancy | truanci | tensely | tens |
| shiftiness | shifti | | |

# Processamento

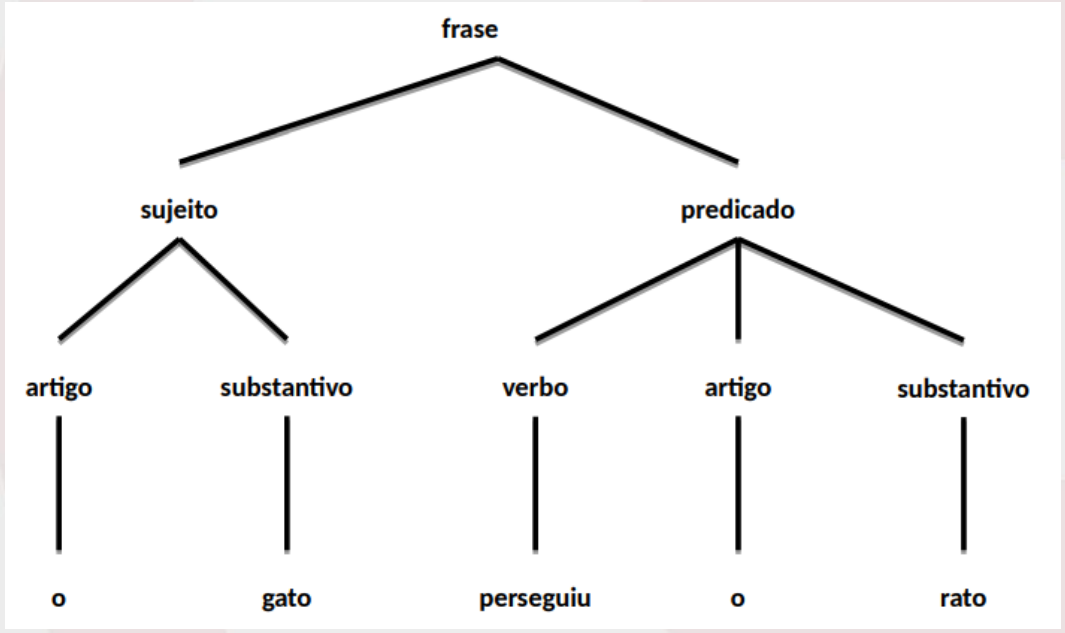
- **Abordagens**:
  - **Simbólica**:
    - Se baseia nas regras linguísticas sem ambiguidades e bem estruturadas.
  - **Estatística**:
    - Utiliza modelos matemáticos e estatísticos.
  - **Conexionista**:
    - Utiliza o aprendizado e teorias de representação do conhecimento.
  - **Híbrida**:
    - Combina as demais abordagens.

# Processamento

- **Fonologia**:
  - Identifica e interpreta os sons que formam as palavras.
- **Morfologia**:
  - Cuida da composição das palavras e sua natureza, dividindo-as em morfemas.
  - Examina os modos que as palavras se desmembram em componentes.
- **Análise sintática (*parsing*)**:
  - Analisa a formação da sentença.
  - Extrair informações de uma frase representada por meio de uma gramática e árvores sintáticas.

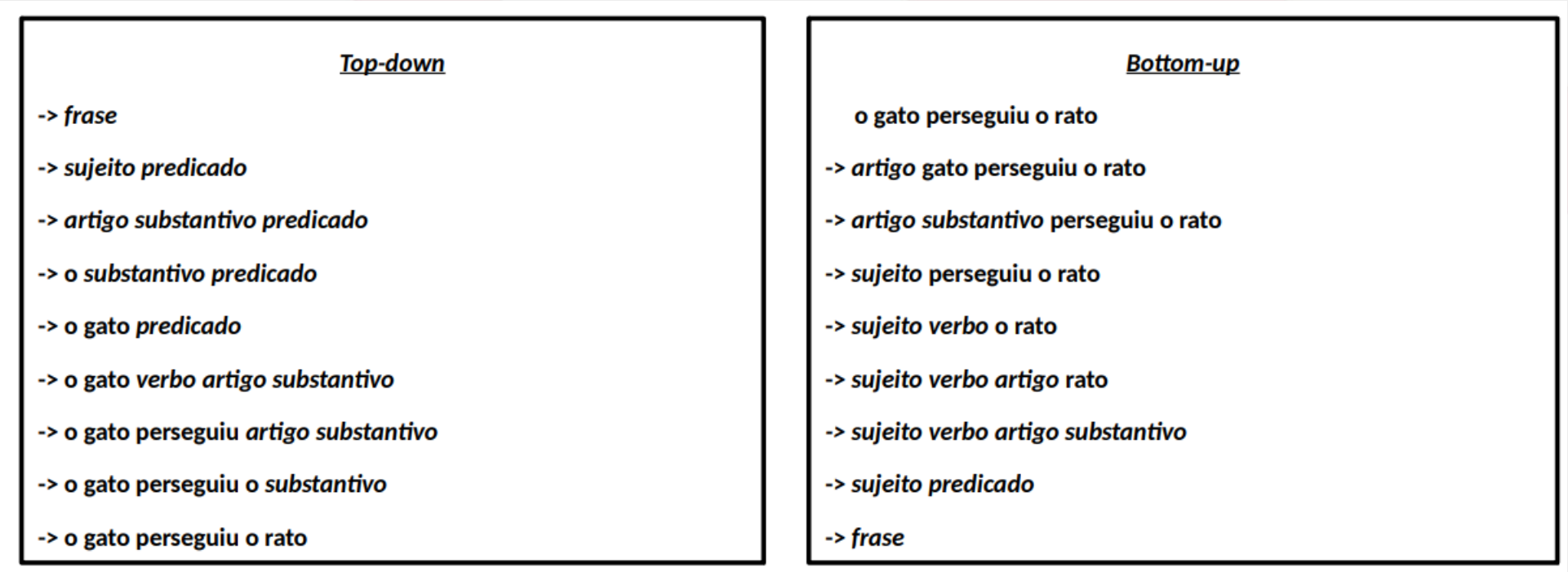
# Processamento

- **Análise sintática (*parsing*)**:
  - Analisa a formação da sentença / a gramática.
  - Extrair informações de uma frase representada por meio de uma gramática e árvores sintáticas.



# Processamento

- **Análise sintática (*parsing*)**:
  - Formas de realização:

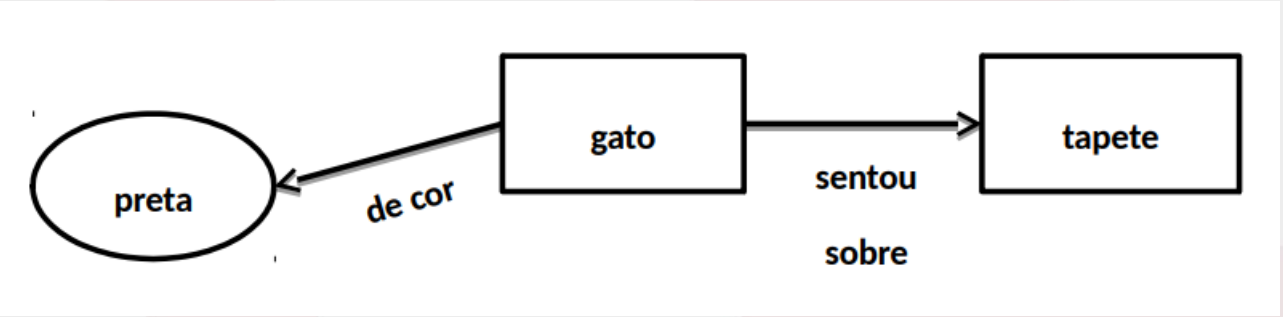


# Processamento

- **Processamento léxico (análise léxica)**:
  - Analisa a entrada e produz uma sequência de símbolos léxicos, captando o significado individual das palavras.
  - Identifica as classes das palavras.
- **Análise semântica**:
  - Se ocupa com o significado da frase extraída da estrutura sintática.
  - Semântica:
    - Considera os significados das palavras.
  - Análise semântica:
    - Análise para extrair um significado de uma declaração.

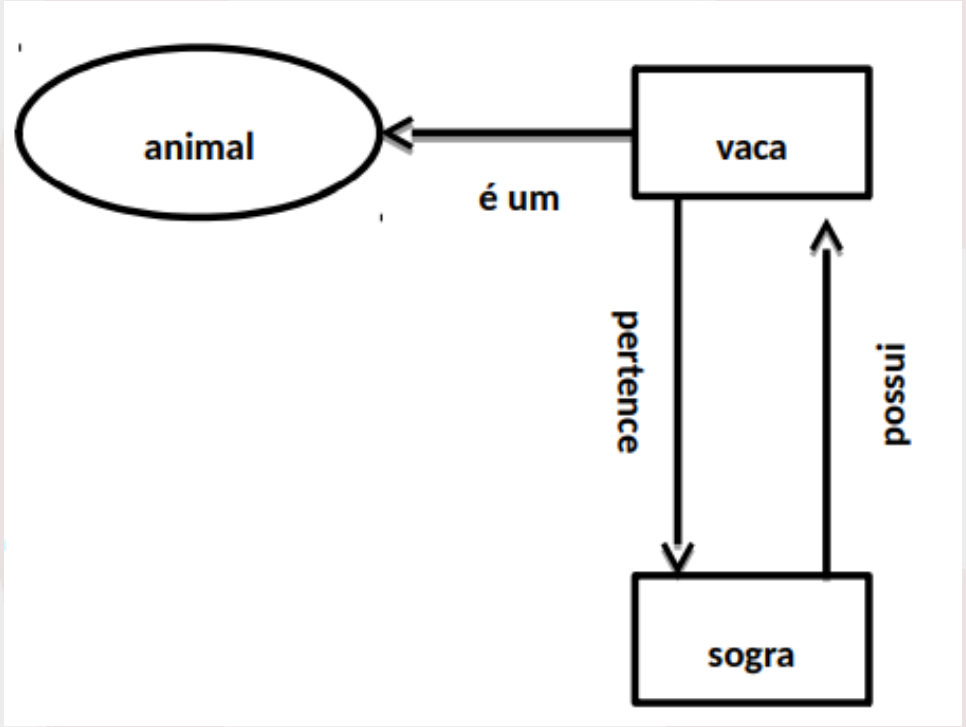
# Processamento

- **Análise semântica:**
  - Exemplo:
    - Frase:
      - O menino comeu o bolo.
    - Semântica:
      - Uma pessoa alimentou/digeriu um alimento.
  - Pode ser representada por uma rede semântica:



# Processamento

- **Análise semântica:**
  - Útil para eliminar ambiguidade:
    - A vaca da minha sogra é branca.



# Processamento

- **Discurso**:
  - Verifica o significado total do texto.
  - Estuda os princípios que governam a produção de sequências estruturadas de frases.
  - Inclui coesão e coerência.
- **Processamento pragmático**:
  - Interpreta os conceitos, averiguando se o significado da análise semântica está correto e ajudando a esclarecê-los.
  - Utilização das palavras em um contexto de interação social.
  - Análise da conversação.