

data

# Sequência de Caracteres e Palavras

2024

# Sumário

- Introdução do Capítulo;
- Conceitos Básicos da Morfologia
  - Morfemas, Token, Type, Lexema, Lexia, Lema, Léxico...
- Processamento Morfológico
  - Sentenciação, Tokenização, Lematização, Pos Tagging
- Ferramentas e Recursos para Processamento Morfológico

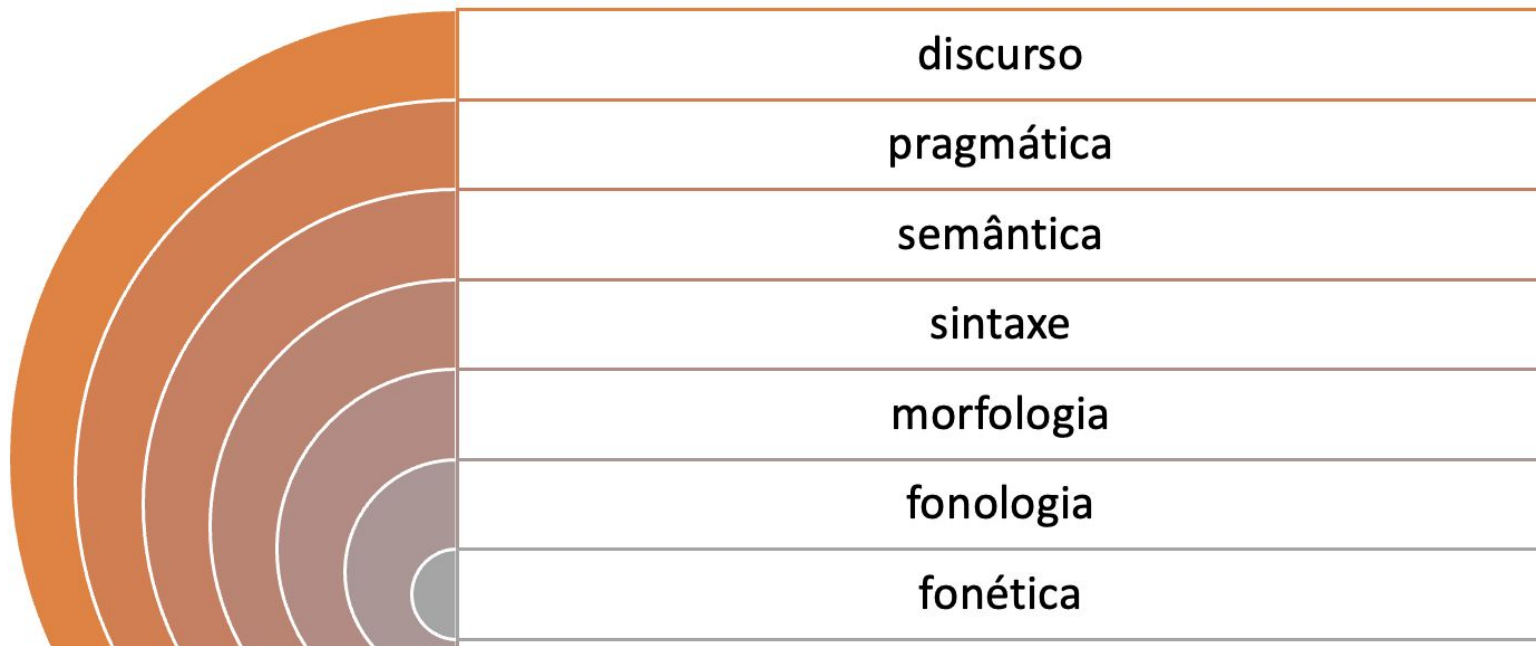




# Introdução



# Sub áreas de Estudo da Linguagem



# Objetivo do Capítulo

- Identificar a unidade mínima quando tratamos, computacionalmente, a língua;
- As subáreas especializadas dos estudos linguísticos entendem como unidade mínima de processamento diferentes elementos;
  - Sabia, Sábia e Sabiá (diferenciação por **fonemas**);
- A morfologia considera o **morfema** como a menor unidade dotada de significado na língua (veremos logo mais);



# Palavras

- O que são
  - Segundo Cabré (1999, p. 20), as palavras são as **unidades de referência** da realidade empregadas pelos falantes;
- São as unidades primárias de processamento em PLN de modo geral;
- Palavra computacional, que se refere a uma unidade linguística que foi criada especificamente para facilitar seu processamento por máquinas;





# Conceitos Básicos de Morfologia



# Morfema

- Unidade mínima significativa;
- Há outras unidades linguísticas que também possuem significado, como a palavra, o sintagma, a frase, a oração, o período, o texto etc;
- Fonema x Morfema;
- Pedacinhos que se juntam para formar as palavras;
- Esses “pedacinhos” podem ser de vários tipos: desinência, raiz, radical, afixo, vogal temática e tema;





# Exemplo

- Palavra “**experimentaria**”:
- “experiment”: significa o conceito lexical de “prova, ensaio, tentativa”;
- “a”: significa que é um verbo da primeira conjugação;
- “ri”: significa que esse verbo está flexionado no tempo futuro do pretérito do modo indicativo;
- “a”: significa que esse verbo está flexionado na terceira pessoa do singular.



# Desinência

- Geralmente ficam no final da palavra;
- Podem marcar gênero e número (no caso dos substantivos e adjetivos) ou marcar número, pessoa, tempo e modo (no caso dos verbos);
- Podem ser classificadas em: nominais ou verbais;
- Exemplo
  - Meninas;
  - Adotasse;



# Raiz/Radical

- Morfema nuclear de uma palavra;
- Constituinte básico que expressa sua base significativa, que designa o significado lexical da palavra.
- É o componente comum a todas as palavras que pertencem à mesma família.
- Exemplo:
  - “menino”, “meninas”, “meninada”, “meninice” e outras possuem a mesma raiz ou radical “menin”.



# Token

- É um termo que significa qualquer sequência de caracteres à qual se atribui um valor.
- Quantidade de palavras + sinais de pontuação = Quantidade de tokens;
- Exemplo:
  - “Eu sempre viajo para Campinas, para Salvador e para Belém.”;
  - Contém 12 tokens, já que, em PLN, os sinais de pontuação (vírgula e ponto final) também são considerados tokens.;



# Type

- Tokens únicos encontrados numa frase ou texto.
- Exemplo anterior possui 10 types (“eu”, “sempre”, “viajo”, “para”, “Campinas”, “,”, “Salvador”, “e”, “Belém” e “.”);



# Léxico

- Léxico corresponde ao conjunto de palavras de uma língua;
- Cada palavra do léxico tem associada a ela uma ou mais triplas com sua categoria gramatical, seu lema e suas características morfológicas;

forma	lema	PoS	atributos morfológicos
sido	ser	AUX	Gender=Masc   Number=Sing   VerbForm=Part
tava	estar	AUX	Abbr=Yes   Mood=Ind   Number=Sing   Person=3   Tense=Imp   VerbForm=Fin
vinha	vir	AUX	Mood=Ind   Number=Sing   Person=3   Tense=Imp   VerbForm=Fin
teríamos	ter	AUX	Mood=Cnd   Number=Plur   Person=1   VerbForm=Fin



# Lexema, Lexia e Lema

- Lexema é sinônimo de unidade lexical, o que implica características de som, forma e significado;
  - “Comprei”;
- Lexia corresponde à realização concreta de um lexema;
  - Em resumo, o lexema é uma representação conceitual enquanto a lexia é a unidade linguística materializada no discurso.;
- Lema é a representação das propriedades sintático-semânticas de um item lexical;
  - “comprar”



# Gramática

- Conjunto de regras e definições que nos permitem escrever de forma padronizada;
  - Em um discurso normal, não poético, a frase “ele leu o livro” e não “o leu ele livro” é correta gramaticalmente.
- Categorias gramaticais: substantivos, adjetivos, nomes próprios, numerais, pronomes, preposições, conjunções, advérbios e verbos;





# Léxico Comum e Léxico Especializado

- O léxico comum corresponde ao conjunto de palavras de uma língua que não têm um “conceito técnico-científico” bem determinado;
- No léxico especializado a palavra assume um significado específico/especial em relação a um sistema de conceitos específico;
  - “DNA”, “Criança”...;



# Palavras Funcionais e Palavras Lexicais

- As palavras funcionais/gramaticais ficam em uma classe fechada;
- A classe fechada tem um número finito de componentes;
- As palavras lexicais ficam em uma classe aberta;
- A classe aberta, por outro lado, acomoda um número bem maior de componentes;



# Processos de formação de palavras

- A derivação é um processo pelo qual novas palavras são criadas adicionando afixos (prefixos, sufixos, infixos etc.) à raiz ou radical;
  - amigo - amigável;
- A composição é um processo em que novas palavras são formadas combinando duas ou mais palavras independentes, ou dois radicais;
  - “beija-flor”, “girassol”...;



# Morfologia e Morfossintaxe

- A morfologia é o ramo da linguística que se concentra no estudo dos morfemas;
- Ela examina como eles se combinam nos processos de flexão e de formação de palavras;
- Em PLN, a morfologia cuida também da classificação dos atributos morfológicos, tais como os traços de gênero, número, modo, tempo, pessoa, voz, caso, entre outros;



# Morfologia e Morfossintaxe

- A morfossintaxe examina como as escolhas morfológicas afetam a organização das palavras em uma sentença e como essas escolhas influenciam a estrutura sintática;
- Ela categoriza as palavras em diferentes classes de palavras a partir da observação de seus atributos morfológicos;
- Em PLN, as classes de palavras são chamadas de part-of-speech ou PoS;





# Processamento Morfológico em PLN



# Motivação

- Para desenvolver praticamente qualquer aplicação de PLN, é necessário realizar etapas que convencionamos chamar de pré-processamento;
- Algumas tarefas usuais são: segmentação do texto em sentenças, separação de palavras, tokenização em subpalavras, normalização de palavras, entre outras;



spaCy

# Sentenciação

- Segmentação do texto em sentenças;
- No caso do português **escrito**, as técnicas usuais se valem da busca de pontuações delimitadoras como “.”, “!” e “?”;
- O desafio é desambiguar essas ocorrências com outros usos dos mesmos caracteres. Um exemplo disto é o caso das abreviações;





# Sentenciação - Abordagens

- Regras, onde são definidos padrões de fim de sentença;
- Abordagens baseadas em aprendizado de máquina supervisionado, ou seja, modelos computacionais treinados sobre conjuntos anotados;
- Abordagens baseadas em aprendizado de máquina não supervisionado, ou seja, modelos computacionais treinados sobre conjuntos não anotados;



# Sentenciação - Spacy

```
!python -m spacy download pt_core_news_md

import nltk
import spacy

[22] nlp = spacy.load("pt_core_news_md")

exampleSentecizing = "Fui à clínica do Dr. Nilo. Quando cheguei, ele não estava atendendo. Será que só perdi meu tempo? Não sei :("
doc = nlp(exampleSentecizing);
for sent in doc.sents:
    print(sent.text);

Fui à clínica do Dr. Nilo.
Quando cheguei, ele não estava atendendo.
Será que só perdi meu tempo?
Não sei :(
```



# Tokenização

- A separação em unidades linguísticas mínimas;
- No caso do português é feita partindo da separação das palavras através de delimitadores;
- Ainda existem problemas de ambiguidade
  - 10.2, beija-flor;



# Tokenização - Spacy



The image shows a Jupyter Notebook interface. The top part is a code cell with Python code for tokenizing a sentence using Spacy. The bottom part is an output cell showing the tokens of the sentence as a list of strings.

```
exampleTokenizing = "Fui à clínica do Dr. Nilo às 10:00 da manhã. No caminho vi um lindo beija-flor"
doc = nlp(exampleTokenizing);
for token in doc:
    print(token.text);
```

Fui  
à  
clínica  
do  
Dr.  
Nilo  
às  
10:00  
da  
manhã  
.  
No  
caminho  
vi  
um  
lindo  
beija-flor



# Tokenização em Sub Palavras

- Tem por objetivo reduzir o vocabulário de trabalho de um modelo de linguagem a um tamanho finito, mas que possa ser usado para representar textos onde o número de types seja potencialmente infinito;
- Consiste em codificar diretamente algumas palavras mais comuns, como “de”, “fazer”, “são” e “feliz”;
- Palavras mais raras, como “desfazer” ou “felizmente” podem ficar fora do vocabulário de trabalho e serem representadas como combinações de subpalavras, respectivamente: “de” + “s” + “fazer” e “feliz” + “mente”;



# Normalização

- Converte as palavras para alguma forma padrão;
- Conversão de versões abreviadas de palavras (e.g., conversão de “vc” para “você”);
- Conversão para caracteres minúsculos (e.g., convertendo “Você” para “você”);
- Lematização (e.g., estabelecendo que “somos” é uma conjugação do verbo “ser”);
- Radicalização (e.g., estabelecendo que “retrabalho” tem o radical “trabalho” precedido do prefixo “re”);



# Lematização - Spacy

```
▶ doc = nlp(exampleTokenizing);  
for token in doc:  
    print(token.lemma_);
```

```
⇒ fui  
a o  
clínica  
de o  
Dr.  
Nilo  
a o  
10:00  
de o  
manhã  
.  
em o  
caminho  
ver  
um  
lindo  
beija-flor
```



# Stemming- NLTK

```
[ ] nltk.download('rslp')  
    nltk.download('punkt')
```

```
▶ exampleStemming = "Andei desorientadamente pelas ruas da cidade com aquela menina."  
  stemmer = nltk.stem.rslp.RSLPStemmer()  
  for token in nltk.word_tokenize(exampleStemming):  
    print(stemmer.stem(token))
```

```
↳ and  
  desorientad  
  pel  
  rua  
  da  
  cidad  
  com  
  aquel  
  menin  
  .
```





# PoS Tagging

- Etiketagem morfofossintática envolve a atribuição de etiquetas gramaticais a cada palavra em um texto, com base na sua classe gramatical e em suas características morfológicas;
- Identificar a função sintática e morfológica das palavras em uma sentença;
- Útil em Tradução automática, análise de sentimentos, geração de resumos, entre outras;
- São universais e valem para a grande maioria das línguas ;



# Pos Tagging - Spacy

```
doc = nlp(exampleTokenizing);
for token in doc:
    print((token.text, token.pos_));
```

( 'Fui', 'VERB' )  
( 'à', 'ADP' )  
( 'clínica', 'NOUN' )  
( 'do', 'ADP' )  
( 'Dr.', 'NOUN' )  
( 'Nilo', 'PROPN' )  
( 'às', 'ADP' )  
( '10:00', 'NOUN' )  
( 'da', 'ADP' )  
( 'manhã', 'NOUN' )  
( '.', 'PUNCT' )  
( 'No', 'ADP' )  
( 'caminho', 'NOUN' )  
( 'vi', 'VERB' )  
( 'um', 'DET' )  
( 'lindo', 'ADJ' )  
( 'beija-flor', 'NOUN' )



# Anotação de atributos morfológicos

- Envolve a marcação ou identificação de informações específicas sobre as características gramaticais e morfológicas de palavras em um texto;
- Incluem características como número, gênero, modo, tempo, pessoa e outras informações semelhantes;
- Capturar e codificar informações gramaticais relevantes de maneira estruturada;

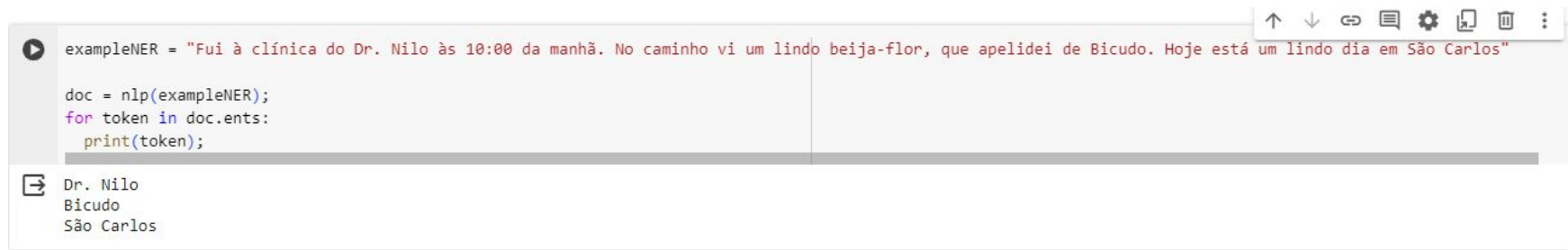


# Outras Tarefas

- Análise sintática automática (tarefa de parsing);
- Segmentação dos constituintes sintáticos dentro da frase (tarefa de chunking);
- Extração ou anotação de entidades nomeadas (tarefa de Named Entity Recognition);



# Outras Tarefas



```
exampleNER = "Fui à clínica do Dr. Nilo às 10:00 da manhã. No caminho vi um lindo beija-flor, que apelidei de Bicudo. Hoje está um lindo dia em São Carlos"
```

```
doc = nlp(exampleNER);  
for token in doc.ents:  
    print(token);
```

Dr. Nilo  
Bicudo  
São Carlos





## Ferramentas e Recursos



# Ferramentas

- Existem ferramentas específicas que podem ser usadas para a análise automática do português;
- [NLTK](#) e [spaCy](#) são módulos mais completos que implementam sub módulos e funcionalidades diversas;
- Outras:
  - [Knowledge Graph](#);



# Recursos

- Corpora anotados com informações morfológicas e morfossintáticas;
- NILC, POeTiSA e LINGUECA são repositórios com recursos como corpora, léxicos e outros recursos lexicais com anotações morfológica e morfossintática;
- PortiLexicon-UD é um recurso que auxilia na tarefa de identificar as unidades de processamento;
- Ele é um léxico para o português que elenca palavras e suas anotações morfossintáticas;





# Recursos

- Utiliza o padrão Universal Dependencies (UD) com etiquetas PoS, lema e etiquetas de atributos morfológicos (gênero, número etc.);
- Na sua versão atual, PortiLexicon-UD possui 1.226.339 entradas;

forma	lema	PoS	atributos morfológicos
bonitas	bonito	ADJ	Gender=Fem   Number=Plur
primeira	primeiro	ADJ	Gender=Fem   Number=Sing   NumType=Ord
cedido	cedido	ADJ	Gender=Masc   Number=Sing   VerbForm=Part
presidente	presidente	NOUN	Number=Sing
quartos	quarto	NOUN	Gender=Masc   Number=Plur



# Considerações Finais

- Abordamos o processamento automático do português no nível da palavra, que é considerado em PLN como a menor unidade de processamento;
- É necessário considerar as pequenas unidades linguísticas que a constituem, que são os morfemas;
- Conceitos da morfologia (morfema, afixo, desinência, radical etc.);
- Conceitos da morfossintaxe (lexema, lexia, léxico, token, type etc.);
- Morfema está para a Morfologia assim como a Palavra está para PLN;



Two L-shaped lines, one blue and one pink, framing the text. The blue line is on the left, and the pink line is on the right.

Obrigado!

