

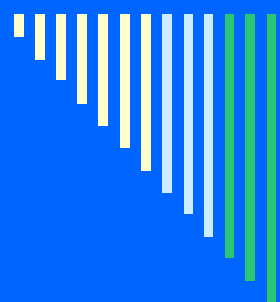
# Corpus comparável: compilação, balanceamento e exploração

Stella E. O. Tagnin

1<sup>a</sup>. Escola Brasileira de Lingüística  
Computacional

**FFLCH – USP – 3 a 5 de setembro de 2007**

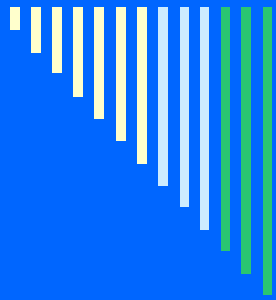
---



# O que é um corpus comparável?

**Textos similares em duas (ou mais) línguas**

- ❑ **Domínio:** Direito, Medicina
- ❑ **Especialidade/Tema:** Direito Contratual; Hipertensão Arterial
- ❑ **Tipologia textual:** artigos científicos, artigos jornalísticos, teses
- ❑ **Período:** semanas, meses, anos
- ❑ **Extensão:** completos, parciais



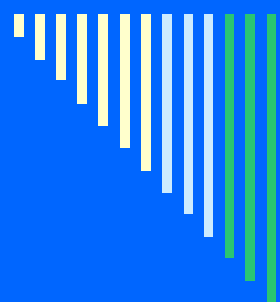
# Como compilar

## Fonte

- Internet** – textos já em formato eletrônico
- Livros** – escanear

## Formatação


- Formato .txt
  - Limpeza: eliminar figuras, tabelas, quadros
  - Inserir cabeçalho
  - Inserir etiquetas, se preciso
- 
- Manter uma cópia no formato original



# Cabeçalho

- quais informações são relevantes para o projeto?
- que outras informações poderiam interessar a outros pesquisadores?
- Reusabilidade

# Cabeçalho

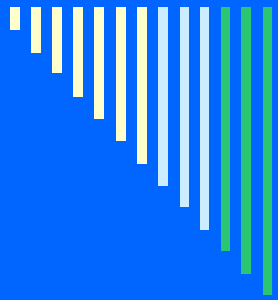


```
<Header>
  <title>
    <filename> </filename>
  </title>
  <author>
    <name></name>
  </author>
  <sourceText>
    <language></language>
    <mode>[mode of delivery of textual
      content]</mode>
    <publisher></publisher>
    <pubPlace>[place of publication]</pubPlace>
    <date></date>
    <copyright>[copyrights holder]</copyright>
  </sourceText>
</Header>
```

# Cabeçalho



```
<Header>
  <title>
    <filename>Adv05red001</filename>
  </title>
  <author>
    <name>Mariazinha da Silva</name>
  </author>
  <sourceText>
    <language>inglês</language>
    <mode>Internet</mode>
    <publisher>Editora Coração</publisher>
    <pubPlace>São Paulo</pubPlace>
    <date>2007</date>
    <copyright>[copyrights holder]</copyright>
  </sourceText>
</Header>
```



# Etiquetagem

- ❑ **morfoossintática (POS-tagging)**
- ❑ **sintática (parsing)**
- ❑ **semântica: campo semântico**
- ❑ **discursiva: ingredientes, modo de fazer, rendimento**
- ❑ **terminológica: termo**



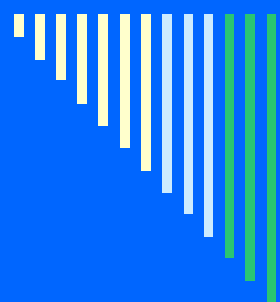
# Etiquetagem morfosintática

- <s>
- Foi\_VAUX
- cercada\_PCP
- de\_PREP | +
- o\_ART
- maior\_ADJ
- sigilo\_N
- a\_ART
- chegada\_N
- de\_PREP | +
- a\_ART
- agência=de=publicidade\_N
- Saatchi\_NPROP
- \$&\_NPROP
- Saatchi\_NPROP
- a\_PREP | +
- o\_ART
- Brasil\_NPROP
- .\_.
- </s>



# Etiquetagem semântica

**For the soup**, preheat the oven to 160°C (350°F / moderate / Gas 4). **<cut>Cut</cut>** **<veg>tomatoes</veg>** lengthwise, discard seeds, place in a medium heatproof dish with **<season>garlic</season>**, olive oil, **<season>salt</season>**, **<season>pepper</season>**, and **<herb>parsley</herb>** and **<herb>basil</herb>** sprigs tied by the stems. **<cook>Bake</cook>** for approximately 1 hour, until **<veg>tomatoes</veg>** are soft and fragrant, let cool and refrigerate for 2 hours, or up to 2 days. **Discard** wilted herbs and blistered tomato skin and puree in a **<appl>blender</appl>** until a smooth paste is obtained (if you want a soup with a more delicate texture, press mixture through a sieve). **Complete** with cold water as to obtain 1 L (1 qt) of soup, adjust **<season>salt</season>** and **<season>pepper</season>**, correct the acidity by adding a pinch of **<season>sugar</season>**, and refrigerate for at least 1 hour, or overnight.



# Etiquetagem semântica/terminológica

Caponata (1 hour and 30  
minutes)

1 onion

2 <term>celery stalks</term>

1 <term>red bell pepper  
</term>

4 fully ripe tomatoes, peeled  
and seeded

1 small deep green zucchini  
(courgette)

2 medium eggplants  
(aubergines)

2 tablespoons <term>pine  
nuts</term>

2 garlic cloves, <term>finely  
chopped</term>

1 <term>bay leaf</term>

1 teaspoon oregano

1/4 cup <term>red wine  
vinegar</term>

1 tablespoon sugar

2 tablespoons capers

2 tablespoons <term>dark  
raisins</term>

1/2 cup slivered green olives

1 cup flat-leaf parsley leaves

1/2 cup basil leaves

olive oil

salt and black pepper <term>to  
taste</term>



# Etiquetagem discursiva

**<titRec>** Pudim de Leite Condensado **</titRec>**

**<coment>** Pudim de leite condensado é uma sobremesa que dispensa elogios, ou qualquer palavra para defini-lo. É simplesmente o máximo!! **</coment>**

**<ingr>** Ingredientes:

1 lata de leite condensado

1 lata de leite

3 ovos

essência de baunilha

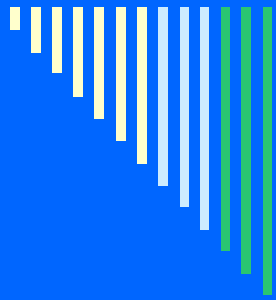
3 colheres de açúcar **</ingr>**

**<modFaz>** Modo de Preparo:

Coloque o açúcar numa forma própria para pudim e leve ao fogo brando para caramelizar a forma. Bater todos os outros ingredientes no liquidificador. Despeje o conteúdo na forma caramelizada. Levar ao forno em banho-maria. **</modFaz>**

**<coment>** Dica: para verificar se o pudim está pronto, fure o pudim com um palito de dente, se o palito sair limpo, é que está pronto, espere esfriar, desenforme e sirva. **</coment>**

---



# Exploração

## Ferramentas

- ❑ Online (integrada): [CorTec](#)
- ❑ Separada (stand-alone): **WordSmith Tools**

## Principais recursos

- ❑ Lista de palavras: frequência, alfabética
- ❑ Concordanciador
- ❑ N-gramas (clusters)