

Instituto Federal do Espírito Santo  
Bacharelado em Sistemas de Informação

Sistema *web* de arquitetura modular para  
processamento de corpora

Ádler Oliveira Silva Neves

Orientador: Me. Ernani Leite Ribeiro Filho

# Sumário

- 1 Introdução ao tópico
- 2 Objetivos
  - Geral
  - Específicos
- 3 Desenvolvimento
  - Um *overview* sobre os objetivos
  - Aprofundando nos objetivos
- 4 Resultados obtidos
  - Treino do etiquetador
  - O sistema desenvolvido
- 5 Conclusão
  - Trabalhos futuros

# O título do trabalho

- Sistema (“elementos que interagem para realizar objetivos”<sup>1</sup>)
  - *web* (navegadores, cliente-servidor, HTML, HTTP)
- de arquitetura (como os elementos estão organizados)
  - modular (baixo acoplamento, *plug-ins*)
- para processamento de (o objetivo é processar *algo*)
  - *corpora* (entrada do sistema)

---

<sup>1</sup>STAIR, R. M.; REYNOLDS, G. W. **Princípios de Sistemas de Informação**. 9ª ed. São Paulo: Cengage Learning, 2011. ISBN: 978-85-221-0797-1. p. 7.

# Corpus

plural: corpora

- O que é?  
Um conjunto de textos relevante para estudar um determinado grupo de pessoas ou um ramo do conhecimento<sup>2</sup>.
- Uma aplicação?
  - Compilação das palavras mais frequentes da língua inglesa em 1921;
  - Revolução no ensino de inglês enquanto língua estrangeira.
- De qual área do conhecimento estamos falando?
  - Processamento automático da linguagem natural;
  - Linguística de corpus.

---

<sup>2</sup>TAGNIN, S. E. **Glossário de linguística de corpus**. São Paulo: HUB Editorial, 2010. p. 358.

## Processamento automático da linguagem natural

Tratamento computacional das estruturas da língua que se repetem.

## Linguística de corpus

O estudo da língua a partir de seus usos em conjuntos de documentos que representam a área estudada.

## Processamento automático de corpora

- É o que o WordSmith faz;
- É o que o CorpusEye faz;
- É o que o SketchEngine faz;
- É o que o Unitex/Gramlab faz;
- É o que o COCA Online Corpus faz;
- É o que o sistema proposto pelo título fará:
  - O que ele faz que os outros não fazem?

**Público alvo:** pesquisadores linguistas, professores de letras, alunos de línguas ou tradutores que não programam.

# O que falta nos atuais?

| Software           | Interface | Gratuito | Corpus<br>fornecido<br>pelo usuário | Tokenizador | Etiquetador | Concordanciador |
|--------------------|-----------|----------|-------------------------------------|-------------|-------------|-----------------|
| WordSmith 5        | Desktop   | Não      | Sim                                 | Sim         | Não         | Sim             |
| Unitex/GramLab     | Desktop   | Sim      | Sim                                 | Sim         | Francês     | Sim             |
| CorpusEye          | Web       | Sim      | Não                                 | Sim         | Sim         | Sim             |
| COCA Online Corpus | Web       | Sim      | Não                                 | Sim         | Sim         | Sim             |
| NoSketch Engine    | Web       | Sim      | Sim                                 | Sim         | Não         | Sim             |
| Sketch Engine      | Web       | Não      | Sim                                 | Sim         | Sim         | Sim             |
| Corpus Slayer      | Web       | Sim      | Sim                                 | Sim         | Sim         | Sim             |

**Tabela 1:** Tabela comparativa resumida de softwares de Processamento de Linguagem Natural

Fonte: O autor

# Objetivo geral

Desenvolver uma aplicação *web* de código aberto para marcação e busca de partes do discurso em corpora, visando ampliar as funcionalidades em relação a *softwares* similares existentes e com interface amigável ao usuário.



## Objetivos específicos

- Analisar comparativamente os recursos das ferramentas WordSmith, CorpusEye, COCA Online Corpus, Unitex/GramLab e Sketch Engine;
- Desenvolver ou adaptar um módulo extrator de sentenças;
- Desenvolver ou adaptar um módulo extrator de lista de palavras;
- Desenvolver ou adaptar um módulo etiquetador de partes do discurso que atue sobre sentenças;
- Desenvolver ou adaptar um módulo concordanciador que suporte busca por etiquetas;
- Integrar os módulos desenvolvidos ou adaptados numa aplicação web;
- Disponibilizar uma ferramenta livre para uso educacional.

# Análise comparativa

- Tabela 4, seção 3.1; (p. 29)
- Várias funcionalidades desejáveis:
  - Não há tempo hábil para implementar todas;
- A tabela 1 era uma versão resumida desta;
- Gerou os objetivos subsequentes.

Um *overview* sobre os objetivos

# Separador de sentenças

Separador de frases

O que faz: Separa um texto em frases;

Desafios: Siglas, abreviações e abreviaturas; (Sr., Sra., V.Exa.)

Implementação: Adaptada do Unitex/Gramlab;





# Concordanciador

O que faz: “extraí todas as ocorrências de uma palavra de busca num corpus juntamente com seu cotexto [...]”<sup>4</sup>;

Implementação: Própria.

---

<sup>4</sup>TAGNIN, S. E. **Glossário de linguística de corpus**. São Paulo: HUB Editorial, 2010. p. 358.

# Tecnologias escolhidas

- Django 1.11;
- Bootstrap 4.0b1;
- FontAwesome 4.7.0;

Na época que o projeto iniciou, estas eram as versões mais recentes.

# Arquitetura modular

- Eventos com tipo:
  - Provedor;
  - Busca;
  - Filtro;
  - Ação.
- Desacoplamento entre partes;
- Módulos substituíveis;
- Uma tela mostrando todas as ações disponíveis sobre um corpus.





# Treino do etiquetador

1 de 3

**Corpus:** Floresta Sintática<sup>5</sup> (ordem de milhão de amostras)

**Etiquetador:** *Unitex/GramLab*

**Problema:** Qual o significado das etiquetas de saída?

- Documentação incompleta;
- Dos 5 artigos citados, apenas um era de acesso público e não trazia dados sobre o significado das etiquetas.

**Problema:** O resultado obtido se compara a quê? O que seria um resultado ruim?

---

<sup>5</sup>LINGUATECA. Projecto Floresta Sinta(c)tica. 2010.

# Treino do etiquetador

2 de 3

Desenvolvido outro etiquetador, para ser o parâmetro de ruim:

**Corpus:** Floresta Sintática<sup>6</sup> (ordem de milhão de amostras)

**Etiquetador:** *YAS-Tagger*

**Funcionamento:** Tabela associativa de vizinhança para etiqueta;

Resultados inesperados levaram à dúvida: “qual seria o impacto se o corpus fosse uma ordem de grandeza menor?”

---

<sup>6</sup>LINGUATECA. Projecto Floresta Sinta(c)tica. 2010.

# Treino do etiquetador

Corpus: Aires<sup>7</sup> (ordem de centena de milhar de amostras)

Etiquetador: *YAS-Tagger*

---

<sup>7</sup>AIRES, R. V. X. Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil. 154 p. Tese (Doutorado) — Universidade de São Paulo, São Carlos, 2000.

# Treinos dos etiquetadores

- ① Floresta Sintática + Unitex/GramLab
- ② Floresta Sintática + YAS-Tagger
- ③ Aires + YAS-Tagger

# Floresta Sintática + Unisex/GramLab

- Precisão  $\approx 60,76\%$ ;
- Precisão concentrada em 3 das 4 etiquetas mais prevalentes:
  - PREP;
  - PRON;
  - V;
- A prevalência da saída do etiquetador não apresenta uma clara relação com a frequência no treino e teste.

# Floresta Sintática + Unites/GramLab

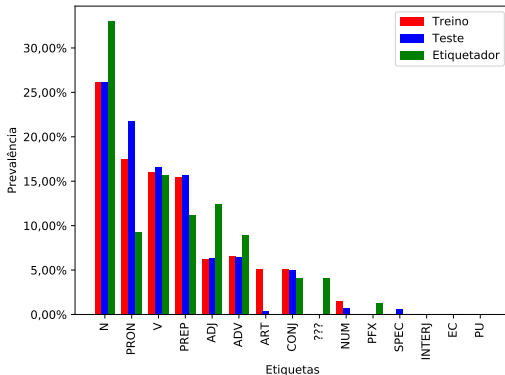


Figura 1: Prevalência das 15 etiquetas mais frequentes do conjunto de treino, teste e etiquetado pelo *Unites/GramLab*

Fonte: O autor





# Floresta Sintática + YAS-Tagger

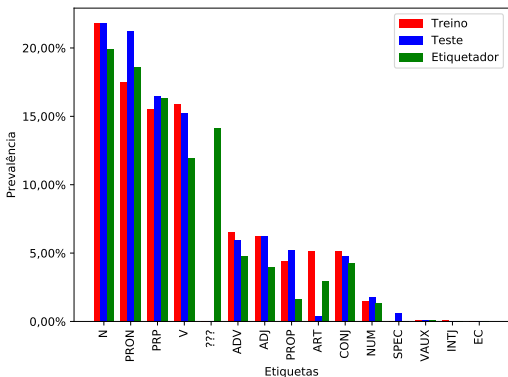


Figura 2: Prevalência das 15 etiquetas mais frequentes do conjunto de treino, teste e etiquetado pelo *YAS-Tagger*

Fonte: O autor



# Aires + YAS-Tagger

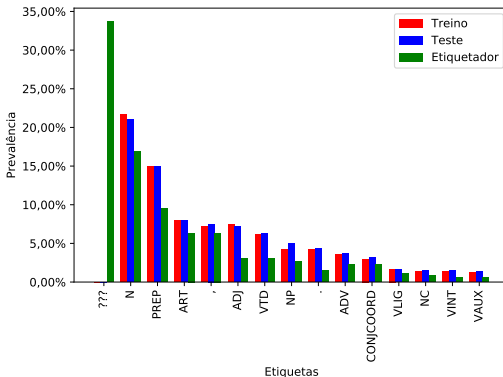


Figura 3: Prevalência das 15 etiquetas mais frequentes do conjunto de treino, teste e etiquetado pelo *YAS-Tagger* sobre o copus de Aires

Fonte: O autor





*Os corpora aparecerão aqui  
depois que você entrar na plataforma*

### Entrar

Usuário

Senha

**Entrar**

Esqueci minha senha

Registrar

Figura 4: Página inicial do sistema desenvolvido

Fonte: O autor

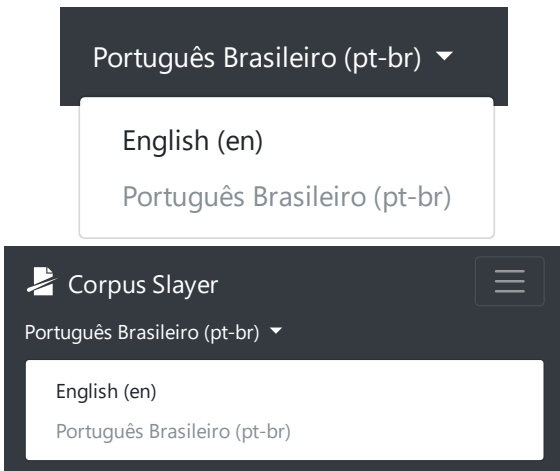


Figura 5: Detalhe do seletor de idiomas, *desktop* em cima e em dispositivos móveis em baixo

Fonte: O autor

The screenshot displays the 'Corpus Slayer' web application interface. At the top, a dark navigation bar contains the application name 'Corpus Slayer', a language dropdown set to 'Português Brasileiro (pt-br)', and links for 'Configurações' and 'Sair'. The main content area features three corpus cards: 'Oriente Médio Grande' (48 documents), 'Fireworks' (1 document), and 'Oriente Médio Pequeno' (1 document). Each card includes a green 'Detalhes' button, creation and modification dates, and a '+ Adicionar corpus' button at the bottom. To the right, a large grey box displays a 'Bem Vindo!' message with the text 'Conectado como adler'. The footer contains the text 'Site esculpido à mão com amor usando Django' and links for 'Ajuda', 'Privacidade', and 'Termos'.

**Figura 6:** Página inicial do sistema desenvolvido, mostrando a listagem dos corpora dum usuário

Fonte: O autor



**Oriente Médio Grande**  
48 documentos

[Detalhes →](#)

Criado 2 meses, 1 semana atrás

Modificado 2 meses, 1 semana atrás

**Fireworks**  
1 documentos

[Detalhes →](#)

Criado 2 meses, 1 semana atrás

Modificado 2 meses, 1 semana atrás

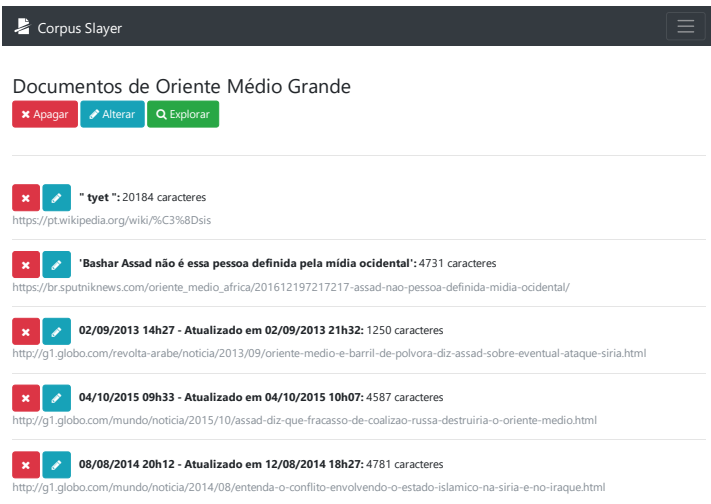
**Oriente Médio Pequeno**  
1 documentos

[Detalhes →](#)

Criado 2 meses, 1 semana atrás

Modificado 2 meses, 1 semana atrás

[+ Adicionar corpus](#)



The screenshot shows the 'Corpus Slayer' application interface. At the top, there is a dark header with the application name and a menu icon. Below the header, the main content area is titled 'Documentos de Oriente Médio Grande'. Underneath the title, there are three action buttons: 'Apagar' (red), 'Alterar' (blue), and 'Explorar' (green). The main area displays a list of five document entries, each with a red 'x' icon, a blue pencil icon, a title, a character count, and a URL. The entries are:

- Document 1: Title: " tyet ": 20184 caracteres; URL: <https://pt.wikipedia.org/wiki/%C3%8Dsis>
- Document 2: Title: 'Bashar Assad não é essa pessoa definida pela mídia ocidental': 4731 caracteres; URL: [https://br.sputniknews.com/oriente\\_medio\\_africa/201612197217217-assad-nao-pessoa-definida-midia-ocidental/](https://br.sputniknews.com/oriente_medio_africa/201612197217217-assad-nao-pessoa-definida-midia-ocidental/)
- Document 3: Title: 02/09/2013 14h27 - Atualizado em 02/09/2013 21h32: 1250 caracteres; URL: <http://g1.globo.com/revolta-arabe/noticia/2013/09/oriente-medio-e-barril-de-polvora-diz-assad-sobre-eventual-ataque-siria.html>
- Document 4: Title: 04/10/2015 09h33 - Atualizado em 04/10/2015 10h07: 4587 caracteres; URL: <http://g1.globo.com/mundo/noticia/2015/10/assad-diz-que-fracasso-de-coalizacao-russa-destruiria-o-oriente-medio.html>
- Document 5: Title: 08/08/2014 20h12 - Atualizado em 12/08/2014 18h27: 4781 caracteres; URL: <http://g1.globo.com/mundo/noticia/2014/08/entenda-o-conflito-envolvendo-o-estado-islamico-na-siria-e-no-iraque.html>

Figura 7: Lista de documentos dum corpus

Fonte: O autor

## Documentos de Oriente Médio Grande

 x Apagar Alterar Explorar

" **tyet** ": 20184 caracteres

<https://pt.wikipedia.org/wiki/%C3%8Dsis>



'**Bashar Assad não é essa pessoa definida pela mídia ocidental**': 4731 caracteres

[https://br.sputniknews.com/oriente\\_medio\\_africa/201612197217217-assad-nao-pessoa-definida-n](https://br.sputniknews.com/oriente_medio_africa/201612197217217-assad-nao-pessoa-definida-n)



Figura 8: Opções de análises para um corpus

Fonte: O autor

## Análise - Oriente Médio Grande

✚ Enviar corpus construído com o BootCaT [bootcat\_upload]

✘ Descartar dados processados [Unitex/GramLab]

✘ Descartar dados processados [MXTERMINATOR]

✘ Descartar dados processados [MXPOST]

✘ Descartar dados processados [TreeTagger]

→ Processar Corpus [Unitex/GramLab]

→ Processar Corpus [MXTERMINATOR]

→ Processar Corpus [MXPOST]

→ Processar Corpus [TreeTagger]

☰ Lista de Sentenças [Unitex/GramLab]

☰ Lista de Sentenças [MXTERMINATOR]

☰ Lista de Sentenças [Unitex/GramLab]

☰ Lista de Sentenças [MXTERMINATOR]

☰ Frequência de Palavras [Unitex/GramLab]

☰ Lista de Palavras [Unitex/GramLab]

☰ Autômato de Sentença [Unitex/GramLab]

☰ Autômato do Texto [Unitex/GramLab]

📁 Texto Etiquetado [MXPOST]

📁 Texto Etiquetado [TreeTagger]

☰ Concordanciador [concord]



## Lista de Sentenças - Oriente Médio Pequeno








1.  Frente dos rebeldes sírios em Aleppo desmorona e leva pânico a refugiados
2.  Ofensiva lançada pelas tropas de Bashar al Assad com apoio russo encurrala refugiados
3.  Recomandar no Facebook
4.  Um guia para entender quem é quem no complexo conflito da Síria
5.  As linhas rebeldes no norte da Síria desmoronam rapidamente nos últimos dias, por causa da intensificação da ofensiva contra Aleppo realizada por forças leais ao presidente Bashar al Assad , apoiadas por combatentes xiitas e pela aviação russa.
6.  Na última segunda-feira, as fileiras do regime se situavam nos arredores de Tal Rifat, a apenas 20 quilômetros da passagem fronteiriça de Öncüpinar/Bab al Salam, entre a Turquia e a Síria.

Figura 9: Lista de sentenças dum corpus

Fonte: O autor

 Corpus Slayer

 Lista de Palavras - Oriente Médio Pequeno

## Palavras não reconhecidas

Afrin Ahrar Alepo Aleppo Ansari Assad Bab Bashar Corps EI EL Facebook Hassan  
Hawa Hazer Idlib jihad jihadistas Kazim Kenyo Kerem Killis Kinik Marea Mercy Nusra  
Nyirjesy Ôncüpinar Qaeda Recomanar Reyhanli Rifat salafista Salam Salih Sham




| Palavra Composta   | Lema      | Gramática & Semântica | Inflexão   |
|--|-----------|-----------------------|--|
|  porta-voz    | porta-voz | N<br>VN               |  |
|  queixaram-se | queixar   | V<br>PRO              |  |

Figura 10: Lista de palavras dum corpus

Fonte: O autor





## Lista de Sentenças Etiquetadas -

### Oriente Médio Pequeno

- 
Frente N dos NPROP rebeldes N sírios N em PREP  
Aleppo N desmorona N e KC leva V pânico ADJ a ART  
refugiados N
- 
Ofensiva ADJ lançada PCP peLas N tropas N de PREP  
Bashar NPROP aL NPROP Assad NPROP com PREP apoio N russo ADJ

Figura 11: Corpus processado pelo *Tree Tagger*

Fonte: O autor



## Concordanciador - Oriente Médio Grande

Corpus etiquetado

TreeTagger

Este campo é obrigatório.

Visibilidade da vizinhança

4 palavras de cada lado

Busca

.ado\_\_VERBO {0,1} terrorista {0,\*} .mic.

Este campo é obrigatório.

Buscar

Decompondo consulta



Referência da notação de busca

Observe a tabela abaixo:

| Busca | Significado       |
|-------|-------------------|
| abc   | A palavra é "abc" |
| \\    | A palavra é "\"   |
| .\.\. | A palavra é "..." |

Figura 12: Tela de busca do concordanciador

Fonte: O autor

# Concordanciador - **Oriente Médio Grande**

Corpus etiquetado

TreeTaqqr

Este campo é obrigatório.

Visibilidade da vizinhança

4 palavras de cada lado

## Busca

Este campo é obrigatório.

## Decompondo consulta

termina com

pula de  até  Palavras

é

pula de  até  Palavras

contém

Referência da notação de busca

Observe a tabela abaixo:

| Busca               | Significado       |
|---------------------|-------------------|
| <code>abc</code>    | A palavra é "abc" |
| <code>\\</code>     | A palavra é "\"   |
| <code>.\.\\.</code> | A palavra é "..." |



## Resultados do Concordanciador - Oriente Médio Grande

- apenas PDEN atua V conforme PREP o ART planejado N . P na PREP  
propaganda N terrorista ADJ gravada PCP com PREP duas NUM câmeras N
- Estado N Islâmico NPROP e NPROP já NPROP considerado PCP o ART  
grupo N terrorista ADJ mais KC perigoso ADJ de PREP todos PRDADJ
- do ADV executivo ADJ turco N e KC considerado PCP um ART  
grupo N terrorista ADJ por PREP Ancara NPROP . P pelos N

Figura 13: Tela de resultados do concordanciador

Fonte: O autor

## Resultados do Concordançador - Oriente Médio Grande

- ← apenas PDEN atua V conforme PREP o ART planejado N , , na PREP  
propaganda N terrorista ADJ gravada PCP com PREP duas NUM câmeras N
- ← Estado N Islâmico NPROP é NPROP já NPROP considerado PCP o ART  
grupo N terrorista ADJ mais KC perigoso ADJ de PREP todos PROADJ
- ← do ADV executivo ADJ turco N e KC considerado PCP um ART  
grupo N terrorista ADJ por PREP Ancara NPROP , , pelos N





 Corpus Slayer ☰

## Documento #88 de Oriente Médio Grande

**Território reivindicado pelo Estado Islâmico no mundo.**  
[https://pt.m.wikipedia.org/wiki/Estado\\_Isl%C3%A2mico\\_do\\_Iraque\\_e\\_do\\_Levante](https://pt.m.wikipedia.org/wiki/Estado_Isl%C3%A2mico_do_Iraque_e_do_Levante)

Território reivindicado pelo Estado Islâmico no mundo.

Desde 2004, a principal meta do grupo é a fundação de um Estado islâmico . [117] [118] O  
 EIII procurou estabelecer-se como um califado, um tipo de Estado islâmico liderado por um

**Figura 14:** Visualização de documento a partir dum clique na seta para trás

Fonte: O autor

 Corpus Slayer ☰

# Documentos encontrados em Orientes Médio Grande

08/08/2014 20h12 - Atualizado em 12/08/2014 18h27  
<http://g1.globo.com/mundo/noticia/2014/08/entenda-o-conflito-envolvendo-o-estado-islamico-na-siria-e-no-iraque.html>

14 pontos-chave sobre o Oriente Médio e o papel do Estado Islâmico  
[http://brasil.elpais.com/brasil/2015/10/11/internacional/1444563614\\_586697.html](http://brasil.elpais.com/brasil/2015/10/11/internacional/1444563614_586697.html)

Figura 15: Listagem de documentos a partir dum clique na seta para trás

Fonte: O autor

- É possível conseguir uma ferramenta comparável às pagas apenas integrando softwares gratuitos existentes;
- Este trabalho deixa uma fonte de inspiração para concordanciadores existentes e futuros uma para aumentar a usabilidade destes por usuários inexperientes;
- Unitex/GramLab tem uma documentação incompleta que cobre apenas o uso da interface.

# Trabalhos futuros

- Garantir que o sistema desenvolvido seja acessível por surdos;
- Adicionar interoperabilidade do sistema desenvolvido com outros sistemas que usam este como execução remota de procedimento;
- Adicionar elementos de rede social, de foma a ser possível compartilhar resultados entre pesquisadores;
- Implementar todos os requisitos levantados, mas não concretizados neste trabalho;
- Implementar pontuador automático para auxiliar pessoas com dislexia.

# Perguntas?