

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**
  - O título do trabalho
  - O domínio da linguística de corpus
- 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
  - modular
- para processamento de
  - corpora

---

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





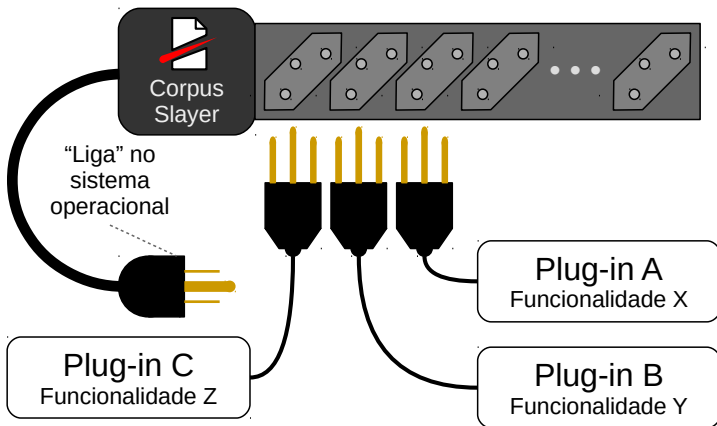


Figura 2: Uma analogia para o conceito de *plug-ins*

Fonte: O autor

# O título do trabalho

- Sistema
  - *web*
- de arquitetura
  - modular
- para processamento de (o objetivo é processar *algo*)
  - corpora (entrada do sistema)





## 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.

# A evolução da linguística de corpus

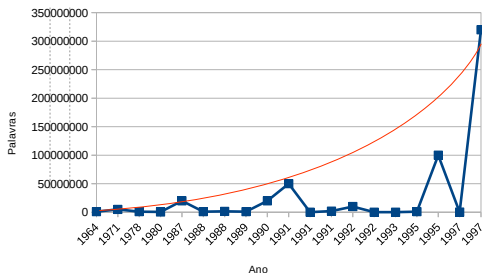


Figura 3: Evolução do tamanho dos corpus no tempo

Fonte: O autor a partir de Sardinha<sup>3</sup>

<sup>3</sup>SARDINHA, Tony Berber. Linguística de Corpus: histórico e problemática. DELTA, São Paulo, v. 16, n. 2, p. 323-367, 2000. Disponível em: <<http://dx.doi.org/10.1590/S0102-44502000000200005>>. Acessado em: 28 jun 2018. p. 330.

O domínio da linguística de corpus

# O linguista antes do computador



**Figura 4:** James Murray e o trabalho que ocupou grande parte de sua vida: *Oxford English Dictionary*

Fonte: Wikipédia<sup>4</sup>

<sup>4</sup><https://en.wikipedia.org/wiki/File:James-Murray.jpg>

## Uma aplicação (que não seja fazer dicionários)

- Compilação das palavras mais frequentes da língua inglesa em 1921;
  - Revolução no ensino de inglês enquanto língua estrangeira.
- 
- Até década de 1960: processamento manual lento, caro e passível de erros;
  - Computadores se popularizaram;
  - 51% do total de domicílios brasileiros de estudantes possuem acesso à internet<sup>5</sup>;
  - E que ferramentas computacionais para processar corpora temos hoje?

---

<sup>5</sup>CETIC.BR. Pesquisa sobre o uso das tecnologias de informação e comunicação nas escolas brasileiras - tic educação 2015. 2016.

# Ferramentas computacionais para linguística de corpus

que fazem o processamento automático da linguagem natural

- É o que o WordSmith faz; (pago)
- É o que o CorpusEye faz; (limitado)
- É o que o SketchEngine faz; (pago)
- É o que o Unitex/Gramlab faz; (limitado)
- É o que o COCA Online Corpus faz; (limitado)
- 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 fornecido 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;
- Deu origem aos 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;







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.  Recomanar 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 Üncüpınar/Bab al Salam, entre a Turquia e a Síria.

Figura 5: Lista de sentenças duma notícia sobre a guerra na Síria<sup>6</sup>

Fonte: O autor

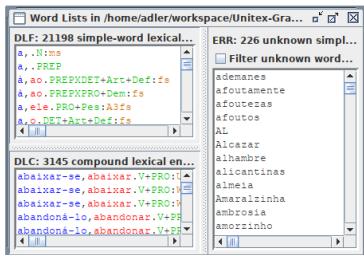
<sup>6</sup>[https://brasil.elpais.com/brasil/2016/02/08/internacional/1454962492\\_021877.html](https://brasil.elpais.com/brasil/2016/02/08/internacional/1454962492_021877.html)

Um *overview* sobre os objetivos

## Extrator de lista de palavras

**O que faz:** Identifica as palavras do texto e as classifica como simples ou composta, seu lema, e suas possíveis flexões;

**Implementação:** Adaptada do Unitex/Gramlab;



**Figura 6:** Lista de palavras do Unitex/GramLab sobre o livro Senhora de José de Alencar

Fonte: O autor

# Concordanciador

O que faz: “extraí todas as ocorrências de uma palavra de busca num corpus juntamente com seu cotexto [...]”<sup>7</sup>;

Implementação: Própria.

---

<sup>7</sup>TAGNIN, S. E. **Glossário de linguística de corpus**. São Paulo: HUB Editorial, 2010. p. 358.

arço a a demissão de Bernard\_Nussbaum , **amigo** de a primeira-dama , de o cargo de assessor jur  
 « Apesar\_de ninguém lá ser mais **amigo** de o rei , a turma continua em a ativa .  
 « Um **amigo** de a família ligou para informar sobre o estado  
 disse que " Zé\_Milionário " é um velho **amigo** seu .  
 , 24 , que assistiu o jogo ao\_lado\_de o **amigo** Brian\_Fagerburg , 24 .  
 laboradores em a campanha\_eleitoral : o **amigo** , empresário e secretário-geral de o PSDB Sérgio  
 « Ronaldo , irmão de Romário , e seu **amigo** Paulo\_Ciro , policial civil , saíram de o condc  
 Cemitério Parque de Animais Jardim de o **Amigo** , em Itapevi ( Grande São\_Paulo ) , tem 20 mil  
 is já foram enterrados em o Jardim de o **Amigo** , todos cães e gatos .  
 as " oferece " o livro a um " distinto **amigo** e colega ´ , até longas mensagens como as de R:  
 no ao\_menos um CD em a brincadeira de " **amigo** secreto " de a empresa em que trabalha .

Figura 7: Concordâncias para a palavra “amigo” gerada pelo CorpusEye

Fonte: O autor

# Etiquetador de partes do discurso

**O que faz:** Atribui etiquetas de partes do discurso a cada palavra da sentença;

**Desafios:** Ambiguidades; (“casa” é verbo ou substantivo?)

**Implementação:** Se o treinamento do Unitex/Gramlab obtiver precisão<sup>8</sup> maior que 75%, será utilizado o etiquetador deste; caso contrário, serão utilizados os etiquetadores treinados por Aires<sup>9</sup>, priorizados por precisão.

---

<sup>8</sup>“denota a proporção de casos preditos como positivos que são considerados positivos reais” POWERS, D. Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness Correlation. **Journal of Machine Learning Technologies**, v. 2, n. 1, p. 37–63, 2011. ISSN 2229-3981. p. 38.

<sup>9</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.

# Etiquetador de partes do discurso

Form	POS sequence #1
Ce	{Ce, ce.DET:ms}
gentleman	{gentleman, .N:ms}
ne	{ne, .XI}
figurait	{figurait, figurer.V:I3s}
dans	{dans, .PREP}
aucun	{aucun, .DET:ms}
comité	{comité, .A:ms}
d'	{de, .PREP}
administration	{administration, .N:fs}
.	{., .PONCT}

Figura 8: Etiquetador do Uitex/GramLab sobre a 9ª sentença do livro A Volta ao Mundo em 80 Dias de Júlio Verne

Um *overview* sobre os objetivos

## Etiquetador de partes do discurso

Form	POS sequence #1
Ce	{Ce, ce, <u>DET</u> :ms}
gentleman	{gentleman, <u>.N</u> :ms}
ne	{ne, <u>.XI</u> }
figurait	{figurait, figurer, <u>.V</u> :I3s}
dans	{dans, <u>.PREP</u> }
aucun	{aucun, <u>.DET</u> :ms}
comité	{comité, <u>.A</u> :ms}
d'	{de, <u>.PREP</u> }
administration	{administration, <u>.N</u> :fs}
.	{., <u>.PONCT</u> }

Figura 9: Etiquetador do Uitex/GramLab sobre a 9ª sentença do livro A Volta ao Mundo em 80 Dias de Júlio Verne

Fonte: O autor



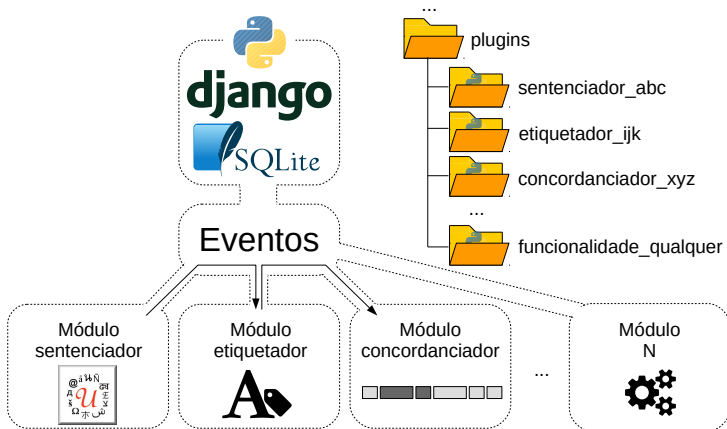


Figura 10: A arquitetura modular utilizada no sistema

Fonte: O autor

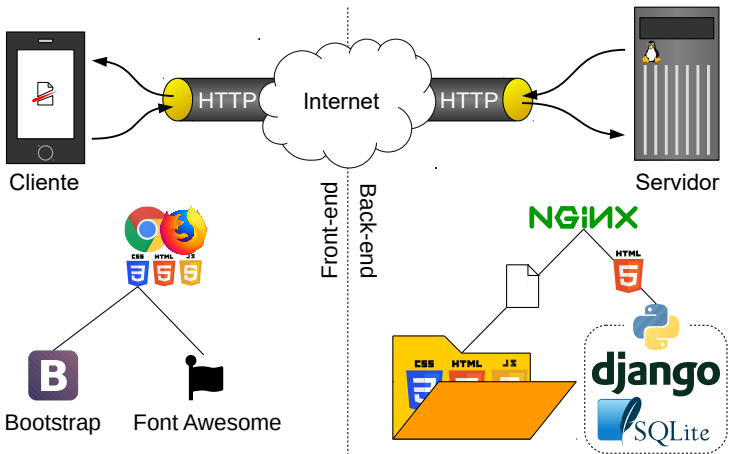


Figura 11: As tecnologias utilizadas nos diferentes espaços do sistema

Fonte: O autor

# Concordanciador

- Problemas a serem resolvidos:

List Chart Collocates Compare KWIC

[POS]

L - - - - - R \*

Keyword in Context (KWIC) Reset

Figura 12: Tela de busca do concordanciador do COCA Online Corpus

Fonte: O autor

- Qual a sintaxe disso?
- Como busco por etiquetas?
- Como o servidor vai entender o que digitei?
- O que eu queria buscar é o que o servidor me retornou?

# Concordanciador

- Ações possíveis:
  - Busca por etiqueta;
  - Busca por palavra exata ou partes desta;
  - Intervalo de fixo ou variável de palavras a ignoradas;
  - Combinação das anteriores.

# Treino do etiquetador

1 de 3

**Corpus:** Floresta Sintática<sup>10</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>10</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>11</sup> (ordem de milhão de amostras)

**Etiquetador:** *YAS-Tagger*

**Funcionamento:** 5 tabelas associativas de trigramas, bigramas e unigramas para etiqueta;

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

---

<sup>11</sup>LINGUATECA. Projecto Floresta Sinta(c)tica. 2010.

# Treino do etiquetador

3 de 3

Corpus: Aires<sup>12</sup> (ordem de centena de milhar de amostras)

Etiquetador: *YAS-Tagger*

---

<sup>12</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 + Unitex/GramLab

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

# Floresta Sintática + UniteX/GramLab

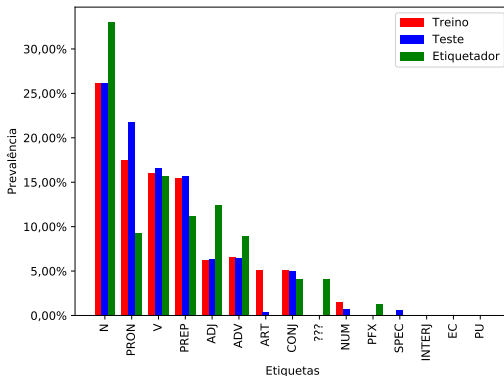


Figura 13: Prevalência das 15 etiquetas mais frequentes do conjunto de treino, teste e etiquetado pelo *UniteX/GramLab*

Fonte: O autor

# Floresta Sintática + YAS-Tagger

- Precisão  $\approx 76,93\%$ ;
- Precisão distribuída mais uniformemente por etiqueta;
- A frequência no conjunto de saída do etiquetador é sempre menor que a frequência desta no treino e teste;

# Floresta Sintática + YAS-Tagger

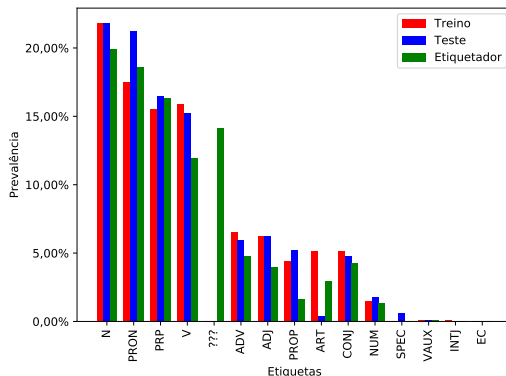


Figura 14: Prevalência das 15 etiquetas mais frequentes do conjunto de treino, teste e etiquetado pelo *YAS-Tagger*

Fonte: O autor

# Aires + YAS-Tagger

- Precisão  $\approx 53,40\%$  (queda de  $23,56\%$ );
- A etiqueta “???” agora representa  $33,74\%$  da saída do etiquetador:
  - aumento de  $138\%$

# Aires + YAS-Tagger

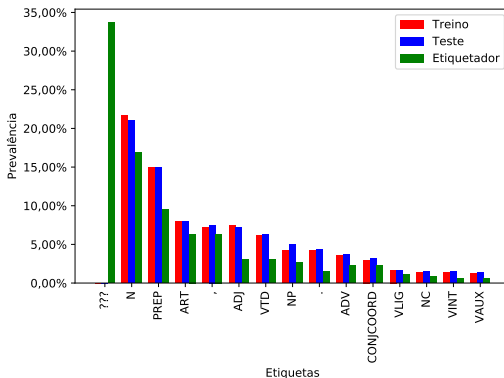


Figura 15: Prevalência das 15 etiquetas mais frequentes do conjunto de treino, teste e etiquetado pelo *YAS-Tagger* sobre o copus de Aires

Fonte: O autor

Etiquetador	Precisão
MXPOST	89,66%
Brill Tagger	88,76%
Tree Tagger	88,47%
YAS-Tagger	76,93%
Unitex/GramLab	60,76%

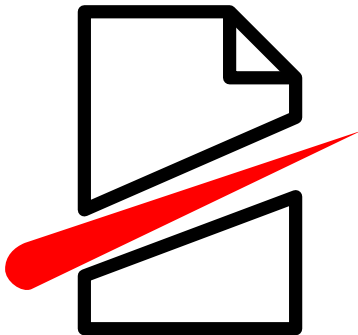
**Tabela 2:** Comparação da precisão entre os etiquetadores *MXPOST*, *Brill Tagger*, *Tree Tagger*, *YAS-Tagger* e *Unitex/GramLab*

Fonte: Aires<sup>13</sup> e o autor.

---

<sup>13</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. p. 82.

# O sistema desenvolvido



<https://corpusslayer.com>



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

Entrar

Usuário

Senha

Entrar

Esqueci minha senha

Registrar

Figura 16: Página inicial do sistema desenvolvido

Fonte: O autor

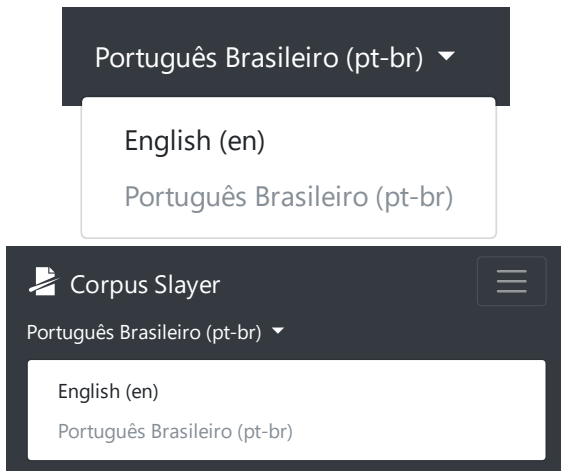


Figura 17: Detalhe do seletor de idiomas, *desktop* em cima e em dispositivos móveis em baixo

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 18: 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

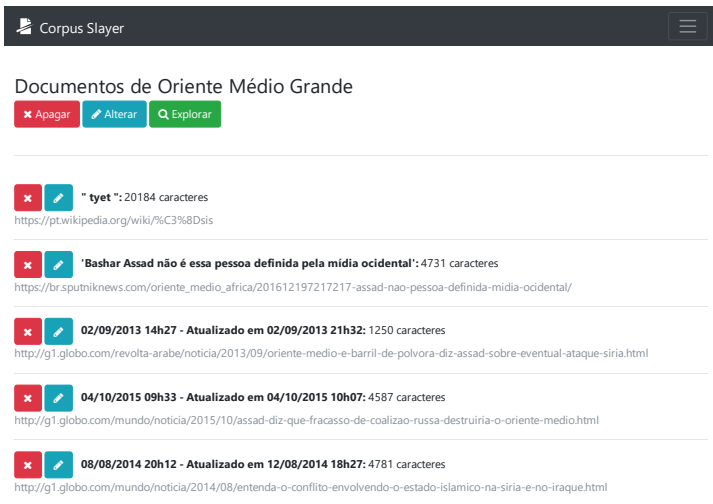
**Oriente Médio Pequeno**  
1 documentos

[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, and a URL. The entries are:

- 1. Title: " tyet ": 20184 caracteres. URL: <https://pt.wikipedia.org/wiki/%C3%8Dsis>
- 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/)
- 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>
- 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>
- 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 19: 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)

The screenshot shows the 'Corpus Slayer' web application interface. At the top, there is a dark header with the application name and a hamburger menu icon. Below the header, the main content area is titled 'Análise - Oriente Médio Grande'. A list of analysis options is displayed, each with a colored bar and a small icon:

- Enviar corpus construído com o BootCaT [bootcat\_upload] (Yellow bar, plus icon)
- Descartar dados processados [Unitev/GramLab] (Red bar, X icon)
- Descartar dados processados [MXTERMINATOR] (Red bar, X icon)
- Descartar dados processados [MPOST] (Red bar, X icon)
- Descartar dados processados [TreeTagger] (Red bar, X icon)
- Processar Corpus [Unitev/GramLab] (Green bar, right arrow icon)
- Processar Corpus [MXTERMINATOR] (Green bar, right arrow icon)
- Processar Corpus [MPOST] (Green bar, right arrow icon)
- Processar Corpus [TreeTagger] (Green bar, right arrow icon)
- Lista de Sentenças [Unitev/GramLab] (Blue bar, list icon)
- Lista de Sentenças [MXTERMINATOR] (Blue bar, list icon)
- Frequência de Palavras [Unitev/GramLab] (Blue bar, list icon)
- Lista de Palavras [Unitev/GramLab] (Blue bar, list icon)
- Autômato de Sentença [Unitev/GramLab] (Blue bar, list icon)
- Autômato do Texto [Unitev/GramLab] (Blue bar, list icon)
- Texto Etiquetado [MPOST] (Blue bar, document icon)
- Texto Etiquetado [TreeTagger] (Blue bar, document icon)
- Concordanciador [concord] (Blue bar, list icon)

At the bottom of the interface, there is a footer with the text 'Site esculpido à mão com amor usando Django' and three links: 'Ajuda', 'Privacidade', and 'Termos'.

Figura 20: 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 Öncüpınar/Bab al Salam, entre a Turquia e a Síria.

Figura 21: 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 Kilis Kinik Marea Mercy Nusra  
Nyirjesy Ôncüpinar Qaeda Recomanar Reyhanli Rifat salafista Salam Salih Sham

## Palavra Composta

## Lema

## Gramática &amp; Semântica

## Inflexão

 porta-voz	porta-voz	<div style="border: 1px solid gray; padding: 2px; display: inline-block;">N</div> <div style="border: 1px solid gray; padding: 2px; display: inline-block;">VN</div>	<div style="border: 1px solid gray; padding: 2px; display: inline-block;">a</div> <div style="border: 1px solid gray; padding: 2px; display: inline-block;">o</div> <div style="border: 1px solid gray; padding: 2px; display: inline-block;">e</div> <div style="border: 1px solid gray; padding: 2px; display: inline-block;">i</div>
 queixaram-se	queixar	<div style="border: 1px solid gray; padding: 2px; display: inline-block;">V</div> <div style="border: 1px solid gray; padding: 2px; display: inline-block;">PRO</div>	<div style="border: 1px solid gray; padding: 2px; display: inline-block;">3</div> <div style="border: 1px solid gray; padding: 2px; display: inline-block;">2</div> <div style="border: 1px solid gray; padding: 2px; display: inline-block;">1</div>

Figura 22: 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 23: 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

termina com (adv) VERBO | pula de (adv) até (adv) Palavras | terrorista (adv)

pula de (adv) até (adv) Palavras | contem (adv)

Referência da notação de busca

Observe a tabela abaixo:

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

Figura 24: 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

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

Este campo é obrigatório.

Buscar

## Decompondo consulta

termina com ado

VERBO

pula de 0 até 1 Palavras

é terrorista

pula de 0 até any Palavras

contém mic

Referência da notação de busca

Observe a tabela abaixo:

<b>Busca</b>	<b>Significado</b>
<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 U 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 25: Tela de resultados do concordanciador

Fonte: O autor

## Resultados do Concordanciador - Oriente Médio Grande

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

2.  Estado N Islâmico NPROP é NPROP já NPROP considerado PCP o ART  
grupo N terrorista ADJ mais KC perigoso ADJ de PREP todos PROADJ



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

Fonte: O autor

 Corpus Slayer

## Documentos encontrados em **Oriente 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 27:** 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 cegos;
- Adicionar interoperabilidade do sistema desenvolvido com outros sistemas que usam este como execução remota de procedimento;
- Adicionar elementos de rede social, de forma 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?