

# SISTEMA WEB DE ARQUITETURA MODULAR PARA PROCESSAMENTO DE CORPORA

# Motivação

- Os uso correto da língua;
- Os benefícios já atingidos;
- Os baixos custos de processamento;
- Crescente presença da internet no cotidiano das pessoas;
- Alguém já deve ter resolvido esse problema...

# Justificativa

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

Fonte: O autor

# Objetivos

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.

- 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.

# Revisão histórica

- Até 1957: Trabalhos manuais
- Chomsky (1957): *Syntactic Structures*
- Década de 1960: Primeiros computadores
- Reconhecimento automático de padrões

## 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.

# SS Medidas estatísticas de avaliação de desempenho de classificadores binários

- Precisão (*precision*)
- Revocação (*recall* ou sensibilidade)
- *F-measure* ( $F_1$  score)
- Acurácia (*accuracy*)
- Taxa de erro (*miss rate*)
- Especificidade (*specificity*)
- Prevalência (*prevalence*)

# Aplicações de internet

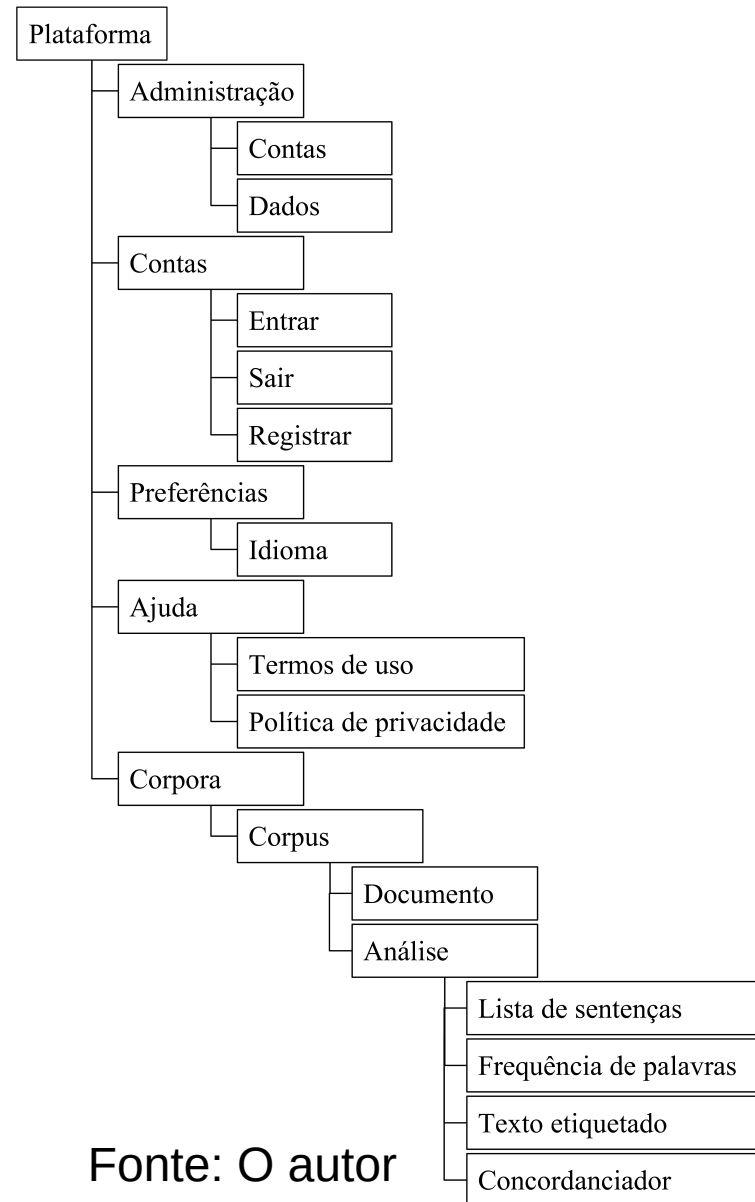
- Lei nº 12.965;
- Terminal burro;
- Arquitetura cliente-servidor;
- Framework de aplicação de internet Django;
- Armazenar com segurança o IP, data e hora de criação de qualquer conteúdo gerado por usuários por no mínimo 6 meses.



# Requisitos levantados

- Possuir extensibilidade através de plug-ins;
- Possuir interface web compatível com tamanhos de tela diversos;
- Processar corpora fornecidos pelo usuário;
- Etiquetar corpora fornecidos pelo usuário;
- Possuir um concordanciador.

# Esboço da navegabilidade



Fonte: O autor

# Fluxo interno de informações

- Eventos com tipo:
  - Provedor;
  - Busca;
  - Filtro;
  - Ação.
- Desacoplamento entre partes;
- Módulos substituíveis;

# Concordanciador

- Problemas a serem resolvidos:
  - Distância entre o usuário e o manual ou referência;
  - Usabilidade.
- Ações possíveis:
  - Busca por etiqueta;
  - Busca por palavra exata ou partes desta;
  - Espaços de número fixo ou variável de palavras a serem aceitas incondicionalmente;
  - Combinação das anteriores.

# Treino do etiquetador

- Corpus: Floresta Sintática
- Etiquetador: Unitex/GramLab
- Qual o significado das etiquetas de saída?

# Treino do etiquetador

- E o resultado obtido seria comparável a quê?
- Para dar a perspectiva de um etiquetador ruim, outro foi escrito baseado em casamento de padrões:
  - O treinamento gera tabelas associativas de tuplas de um, dois ou três palavras e à ela é atribuída a etiqueta mais frequente;
  - A etiquetagem aplica as tabelas sobre o texto, deixando “???” como etiqueta não seja possível atribuir nenhuma.
- Nomeado *YAS-Tagger*.

# Treino do etiquetador

- Dúvida durante a análise dos dados nos resultados: E se trocar o corpus por um menor, como o *YAS-Tagger* se comporta?
- Outro corpus: Aires (2000)

# Treino do etiquetador

- Floresta Sintática + *Unitex/GramLab*
- Floresta Sintática + *YAS-Tagger*
- Aires (2000) + *YAS-Tagger*

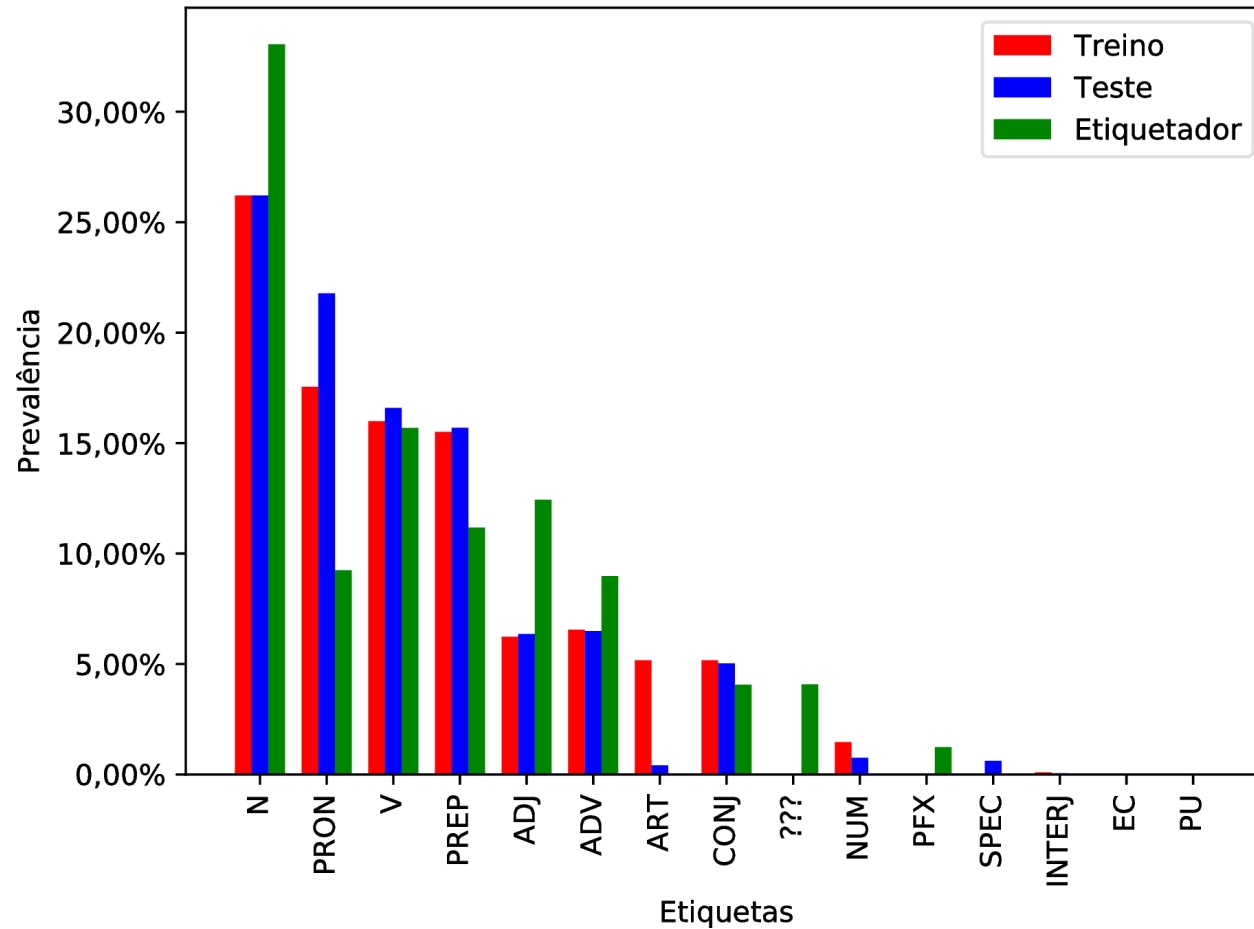


# Resultados

# Floresta Sintática + *Unitex/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 + *Unitex/GramLab*



Fonte: O autor

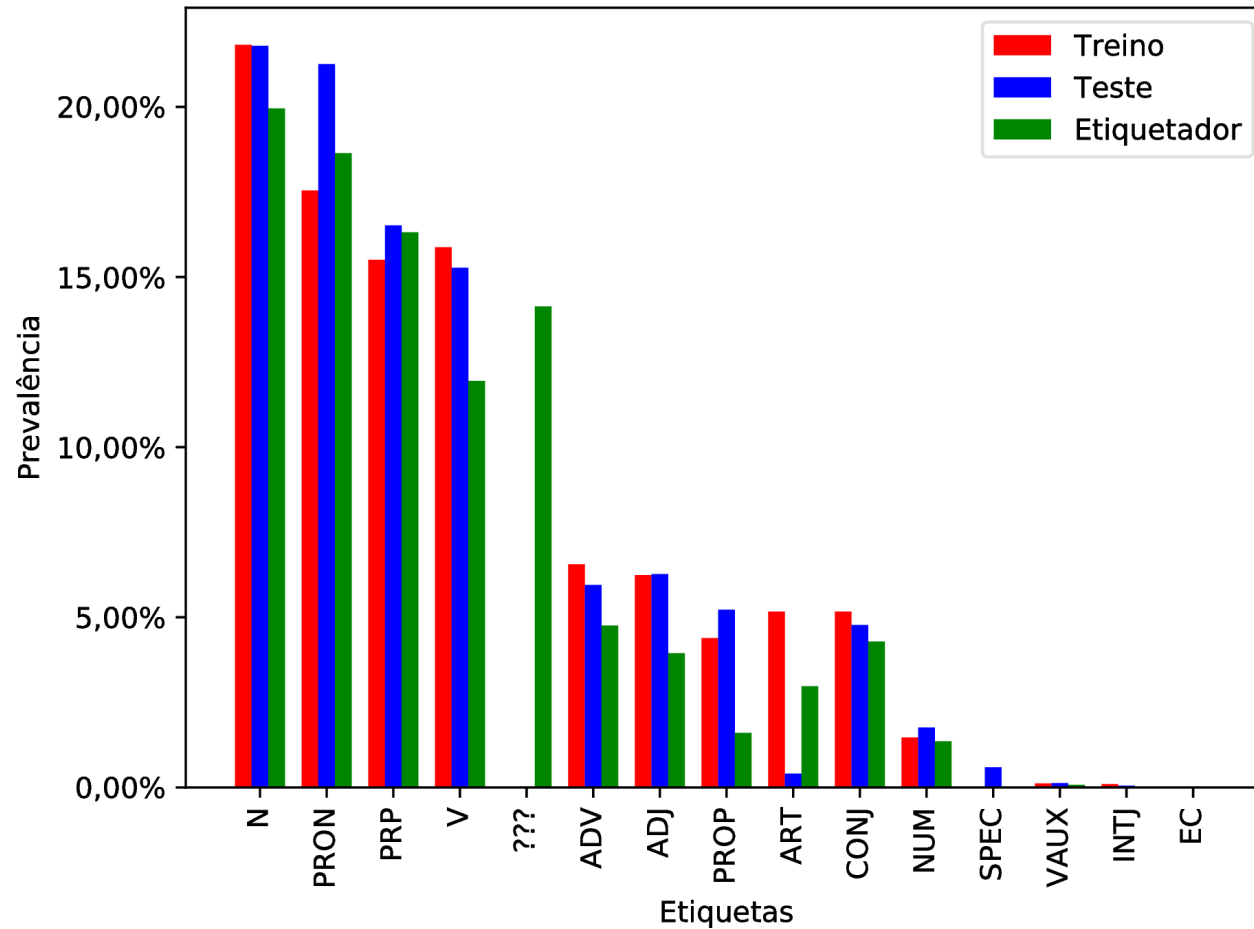
# Floresta Sintática + *Unitex/GramLab*

- Pouco interessante para um servidor, mesmo com fartura de recursos:
  - Mais de 72 horas de processamento;
  - 14GB de uso da memória principal;
  - 30GB de uso de *swap*.

# Floresta Sintática + *YAS-Tagger*

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

# Floresta Sintática + *YAS-Tagger*



# Floresta Sintática + *YAS-Tagger*

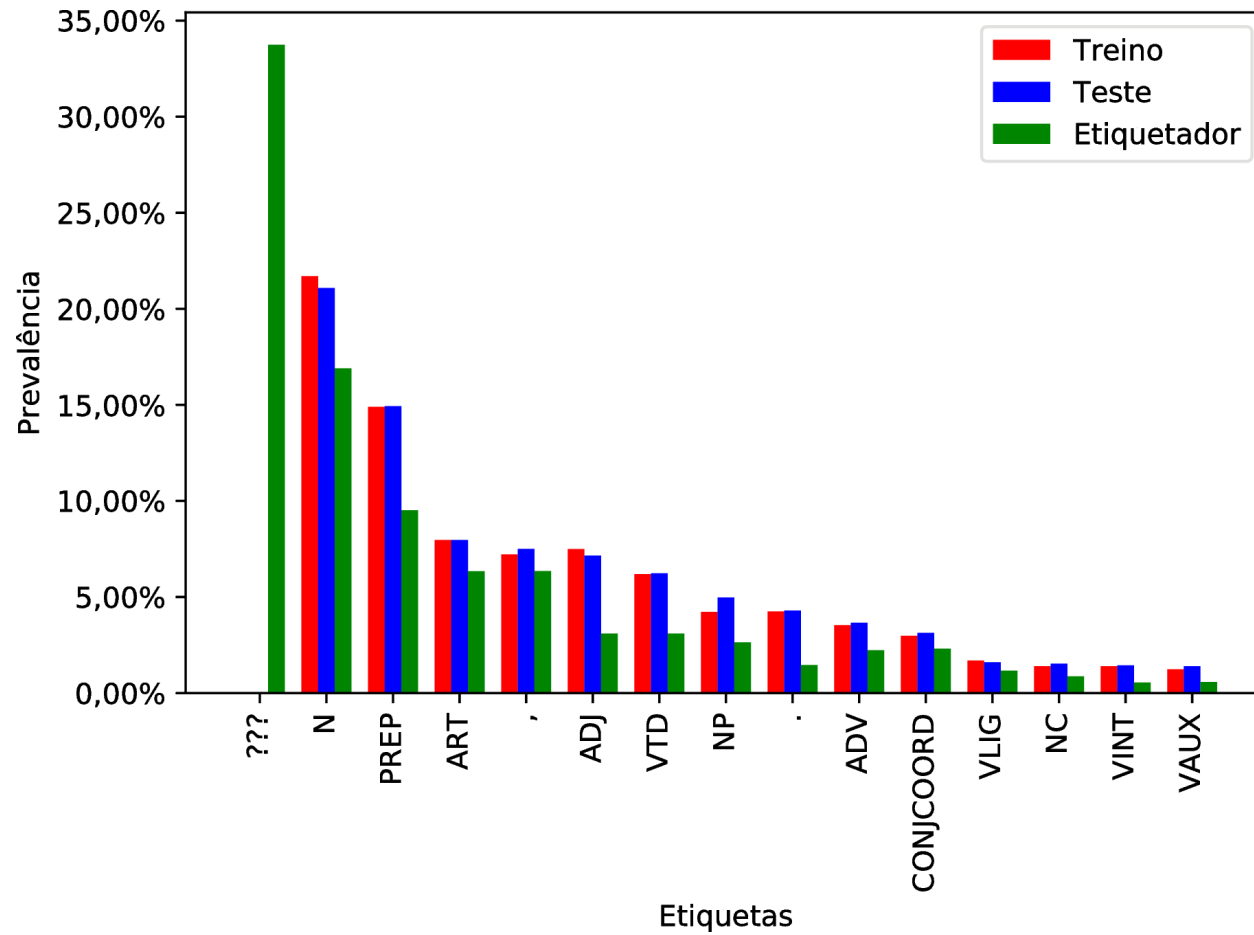
- Quantidade significativa (14,13%) de erros devido à aplicação da etiqueta “???”:
  - E se o corpus fosse menor, qual seria o impacto?

# Aires (2000) + *YAS-Tagger*

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



# Aires (2000) + *YAS-Tagger*



Fonte: O autor

# Aires (2000) + *YAS-Tagger*

- Respondendo à pergunta deixada em aberto:  
“E se o corpus fosse menor, qual seria o impacto?”
  - Catastrófico.

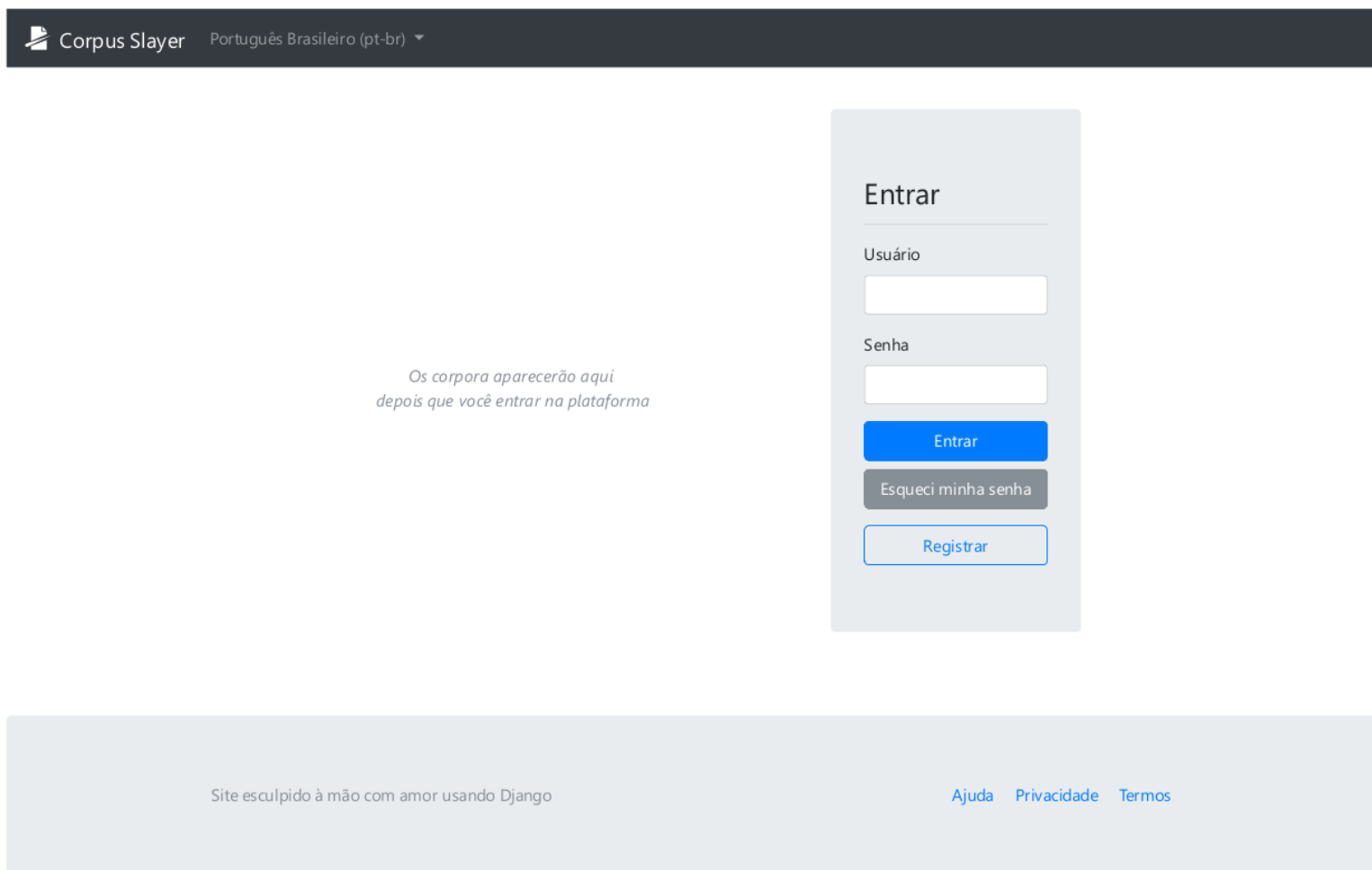
# E qual o “placar”?

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

Fonte: O autor e Aires (2000, p. 82)

- E é essa a ordem de prioridade utilizada para a implementação.

# O sistema desenvolvido



Tela inicial do sistema desenvolvido

Fonte: O autor

# O sistema desenvolvido

Português Brasileiro (pt-br) ▼

English (en)

Português Brasileiro (pt-br)

 Corpus Slayer

Português Brasileiro (pt-br) ▼

English (en)

Português Brasileiro (pt-br)

Seletor de idiomas do sistema desenvolvido

Esquerda: *desktop*

Direita: *mobile*

Fonte: O autor

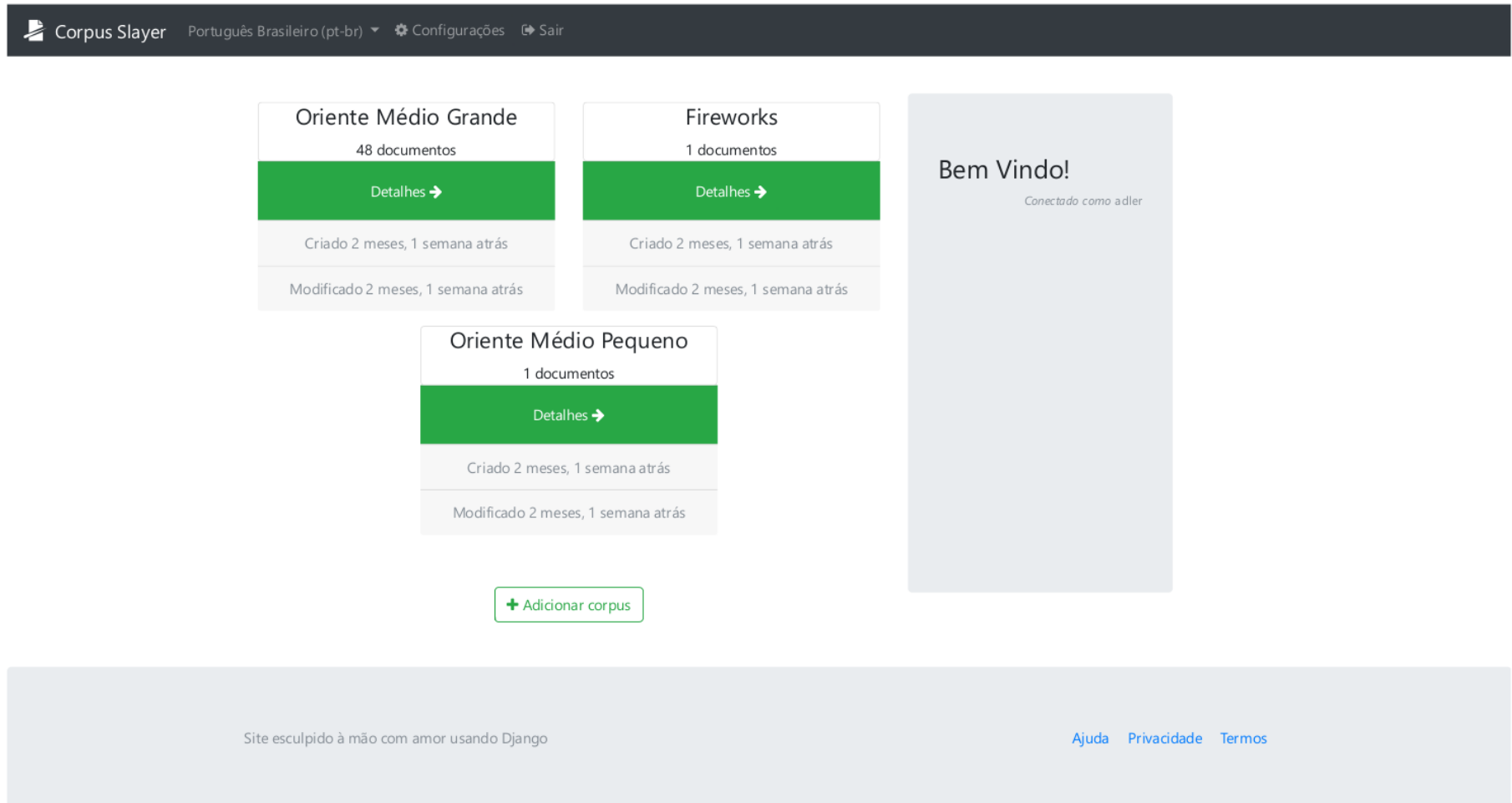
# O sistema desenvolvido



Botões de login do sistema desenvolvido  
Esquerda: *desktop*, canto superior-direito  
Direita: *mobile*, abaixo do seletor de idiomas

Fonte: O autor



# O sistema desenvolvido






Tela inicial do sistema desenvolvido, após autenticação,  
exibindo a lista de corpora

Fonte: O autor



# O sistema desenvolvido

 Corpus Slayer 



## Documentos de Oriente Médio Grande



---

  " **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-midia-ocidental/](https://br.sputniknews.com/oriente_medio_africa/201612197217217-assad-nao-pessoa-definida-midia-ocidental/)



---

  **02/09/2013 14h27 - Atualizado em 02/09/2013 21h32**: 1250 caracteres  
<http://g1.globo.com/revolta-arabe/noticia/2013/09/oriente-medio-e-barril-de-polvora-diz-assad-sobre-eventual-ataque-siria.html>

---

  **04/10/2015 09h33 - Atualizado em 04/10/2015 10h07**: 4587 caracteres  
<http://g1.globo.com/mundo/noticia/2015/10/assad-diz-que-fracasso-de-coalizao-russa-destruiria-o-oriente-medio.html>

---

  **08/08/2014 20h12 - Atualizado em 12/08/2014 18h27**: 4781 caracteres  
<http://g1.globo.com/mundo/noticia/2014/08/entenda-o-conflito-envolvendo-o-estado-islamico-na-siria-e-no-iraque.html>

---

Lista de documentos dum corpus

Fonte: O autor



# O sistema desenvolvido



Opções de análises para um corpus







Fonte: O autor

# Extrator de sentenças e lista de palavras

 Corpus Slayer




## 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.  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 Öncüpinar/Bab al Salam, entre a Turquia e a Síria.

Lista de sentenças dum corpus

Fonte: O autor

# Extractor de sentenças e lista de palavras

 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 Öncüpinar Qaeda Recomanar Reyhanli Rifat salafista Salam Salih Sham

---

Palavra Composta	Lema	Gramática & Semântica	Inflexão
 porta-voz	<input type="text" value="porta-voz"/>	N VN	<span>m s</span> <span>f s</span>
 queixaram-se	<input type="text" value="queixar"/>	V PRO	<span>j 3 p</span>

Lista de palavras dum corpus

Fonte: O autor

# Etiquetador

 Corpus Slayer



Erro enquanto processava o corpus - [Oriente Médio Pequeno](#)

[Retornar às opções de análise](#)

Detalhes:

```
Read 14588 items from ./plugins/nilc_aires/datasets/mxpost/word.voc
Read 240 items from ./plugins/nilc_aires/datasets/mxpost/tag.voc
Read 54206 items from ./plugins/nilc_aires/datasets/mxpost/tagfeatures.contexts
Read 54524 contexts, 116146 numFeatures from ./plugins/nilc_aires/datasets/mxpost/tagfeatures.fmap
Read model ./plugins/nilc_aires/datasets/mxpost/model : numPredictions=240, numParams=116146
Exception in thread "main" java.lang.ArrayIndexOutOfBoundsException: 14588
    at try.<init>()
    at static.c()
    at public.<init>()
    at tagger.TestTagger.main()
```

Erro do *MXPOST* ao etiquetar para o Português Brasileiro utilizado arquivos treinados por Aires (2000) – o que não acontece com a língua inglesa

Fonte: O autor

# Etiquetador

```
$ ./tagger  
YOU MUST RUN THIS PROGRAM IN THE SAME DIRECTORY AS ./start-state-tagger and ./final-state-tagger  
AND ./start-state-tagger and ./final-state-tagger MUST HAVE EXECUTE PERMISSION SET  
$ █
```

Erro do *Brill Tagger* requisitando arquivos com permissão de execução

Fonte: O autor

- Requer permissão de execução para arquivos que receberão dados fornecidos pelo usuário:
  - Possível vulnerabilidade de *Remote Code Execution*



# Etiquetador

 Corpus Slayer



## 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

Corpus processado pelo *Tree Tagger*

Fonte: O autor

# Concordanciador

Corpus Slayer
Português Brasileiro (pt-br) ▾
Configurações
Sair

## Concordanciador - Oriente Médio Grande

Corpus etiquetado

TreeTaqqer ▾

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:

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

Tela de busca do concordanciador

Fonte: O autor

# Concordanciador

## 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

Tela de resultados do concordanciador

Fonte: O autor



# A “seta para trás”



A “seta para trás” presente na tela de resultados do concordanciador, na do etiquetador e nas listas de sentenças e lista de palavras, em destaque

Fonte: O autor

- Leva de volta ao(s) documento(s) que contém o que está em exibição na tela

# A “seta para trás”

 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 FIII procurou estabelecer-se como um califado, um tipo de Estado islâmico liderado por um

Visualização de documento a partir dum clique na “seta para trás”

Fonte: O autor

# A “seta para trás”

 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)

Listagem de documentos a partir dum clique na “seta para trás”

Fonte: O autor

# Conclusão

- 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 forma a ser possível compartilhar resultados entre pesquisadores;
- Implementar todos os requisitos levantados, mas não concretizados neste trabalho.

Dúvidas?