SISTEMA WEB DE ARQUITETURA MODULAR PARA PROCESSAMENTO DE CORPORA

S

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...

S

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

S

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 adptados numa aplicação web.

S

Revisão histórica

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

SS

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*)

SS

Aplicações de internet

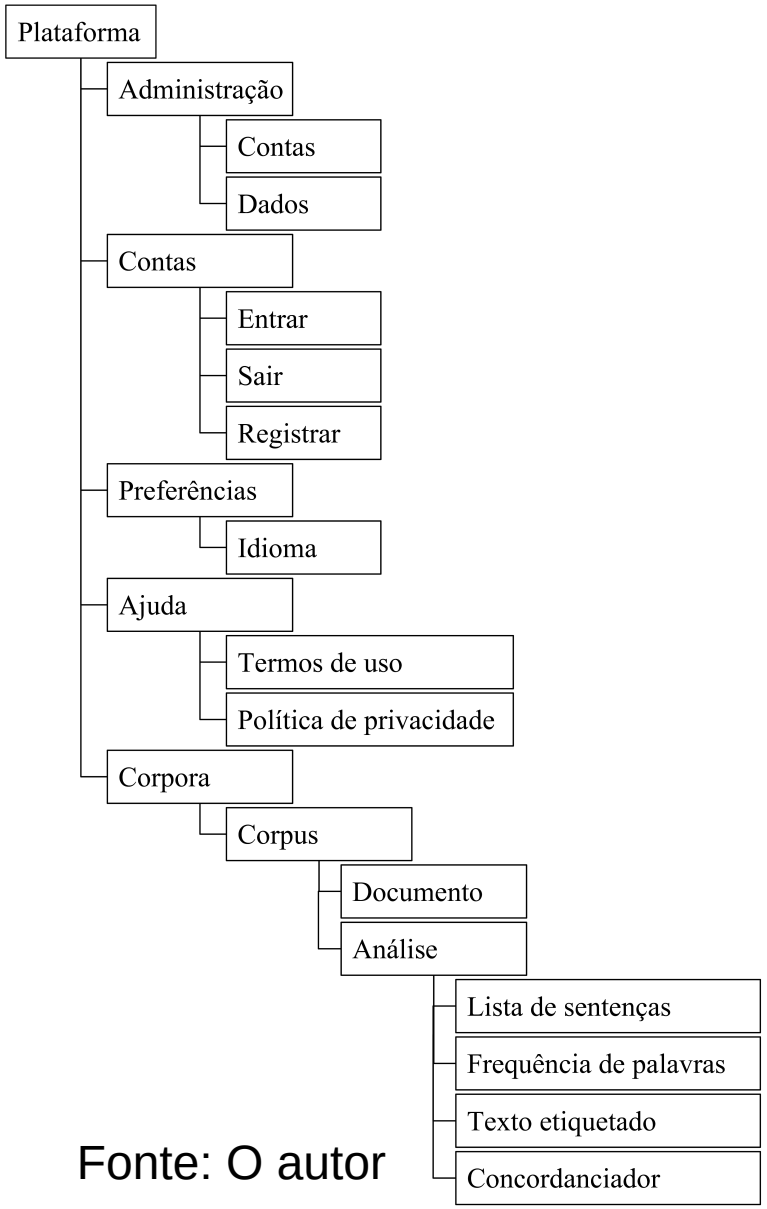
- Lei nº 12.965;
- Terminal burro;
- Arquitetura cliente-servidor;
- Framework de aplicação de internet Django;
- Armazenar com segurança o IP, data e hora de criação de qualquer conteúdo gerado por usuários por no mínimo 6 meses.

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.

S

Esboço da navegabilidade



Fonte: O autor

S

Fluxo interno de informações

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

S

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.

S

Treino do etiquetador

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

S

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.*

S

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)

S

Treino do etiquetador

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

Resultados

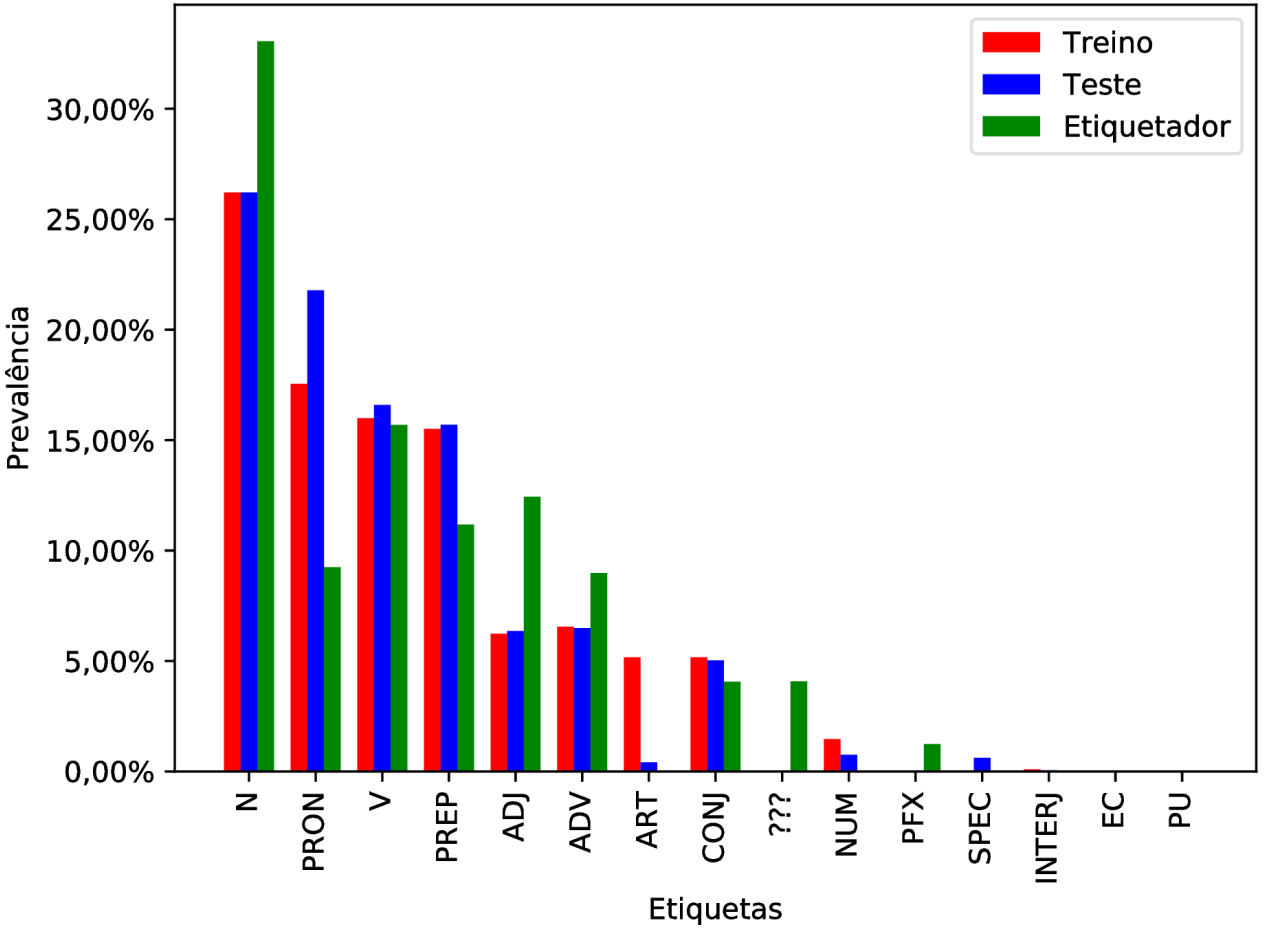
SS

Floresta Sintática + *Unitex/GramLab*

- Precisão ≈ 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.

SS

Floresta Sintática + *Unitex/GramLab*



Fonte: O autor

SS

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.*

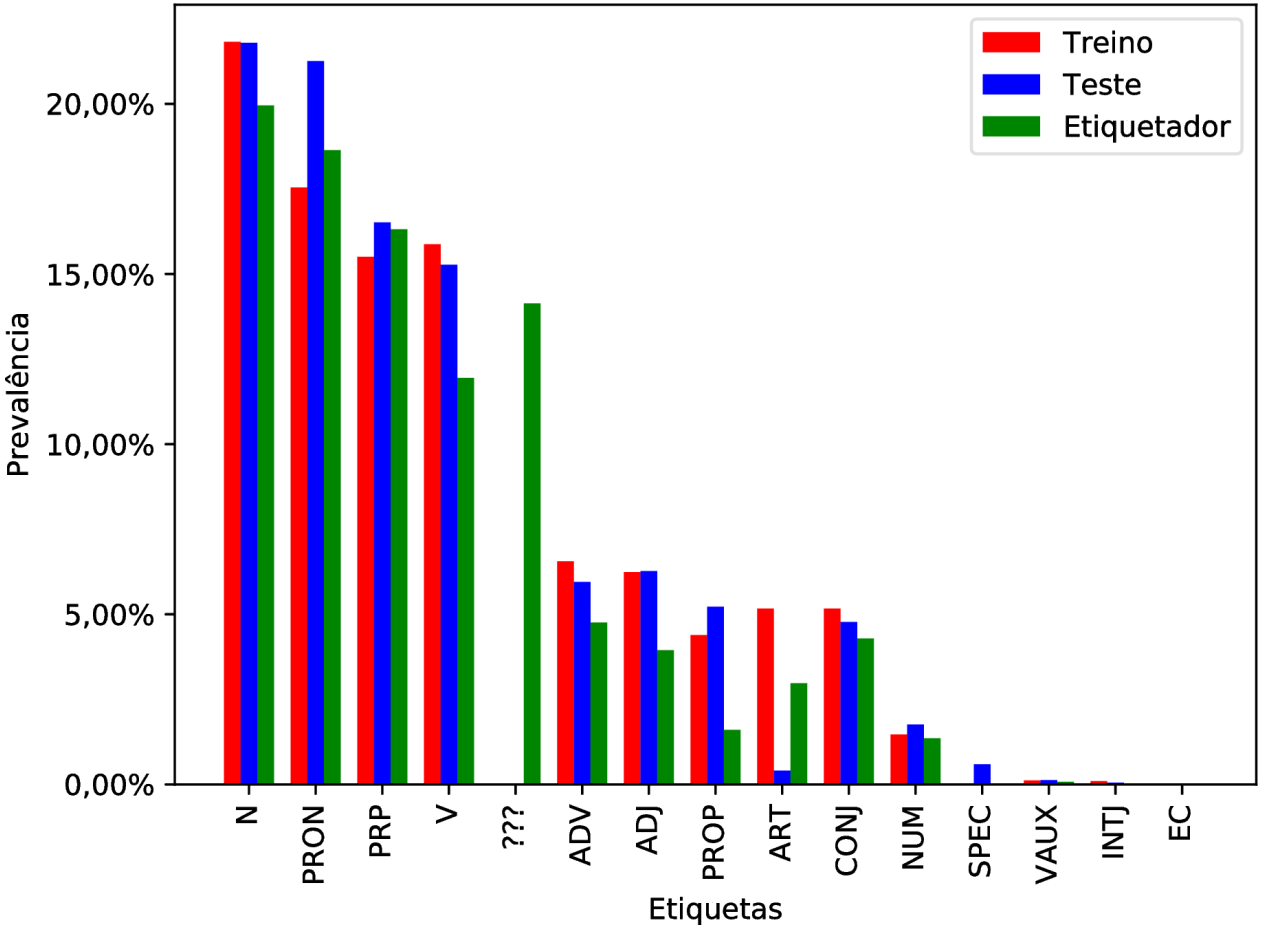
SS

Floresta Sintática + *YAS-Tagger*

- Precisão ≈ 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;

SS

Floresta Sintática + *YAS-Tagger*



Fonte: O autor

SS

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?

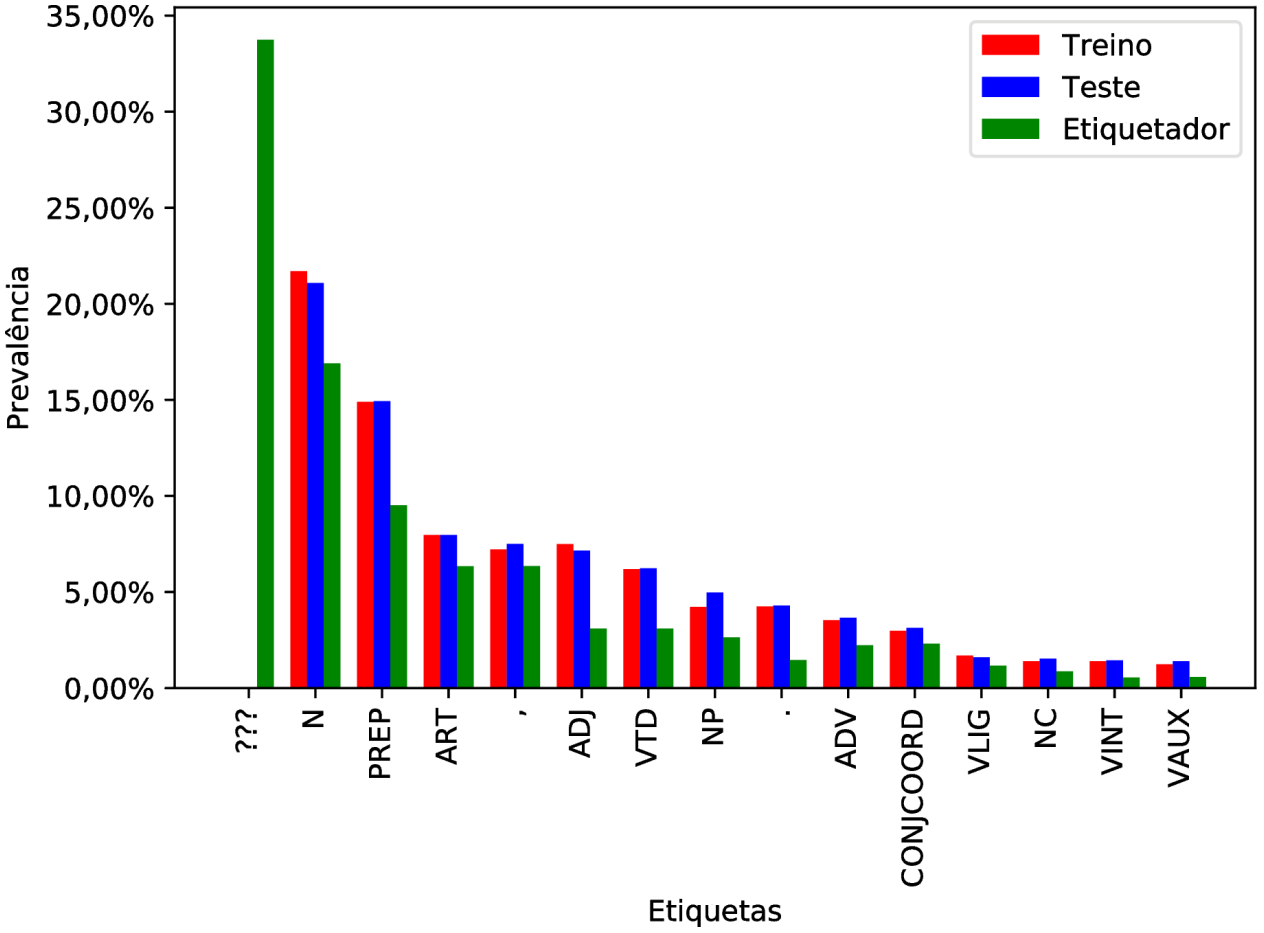
SS

Aires (2000) + *YAS-Tagger*

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

SS

Aires (2000) + *YAS-Tagger*



Fonte: O autor

SS

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”?

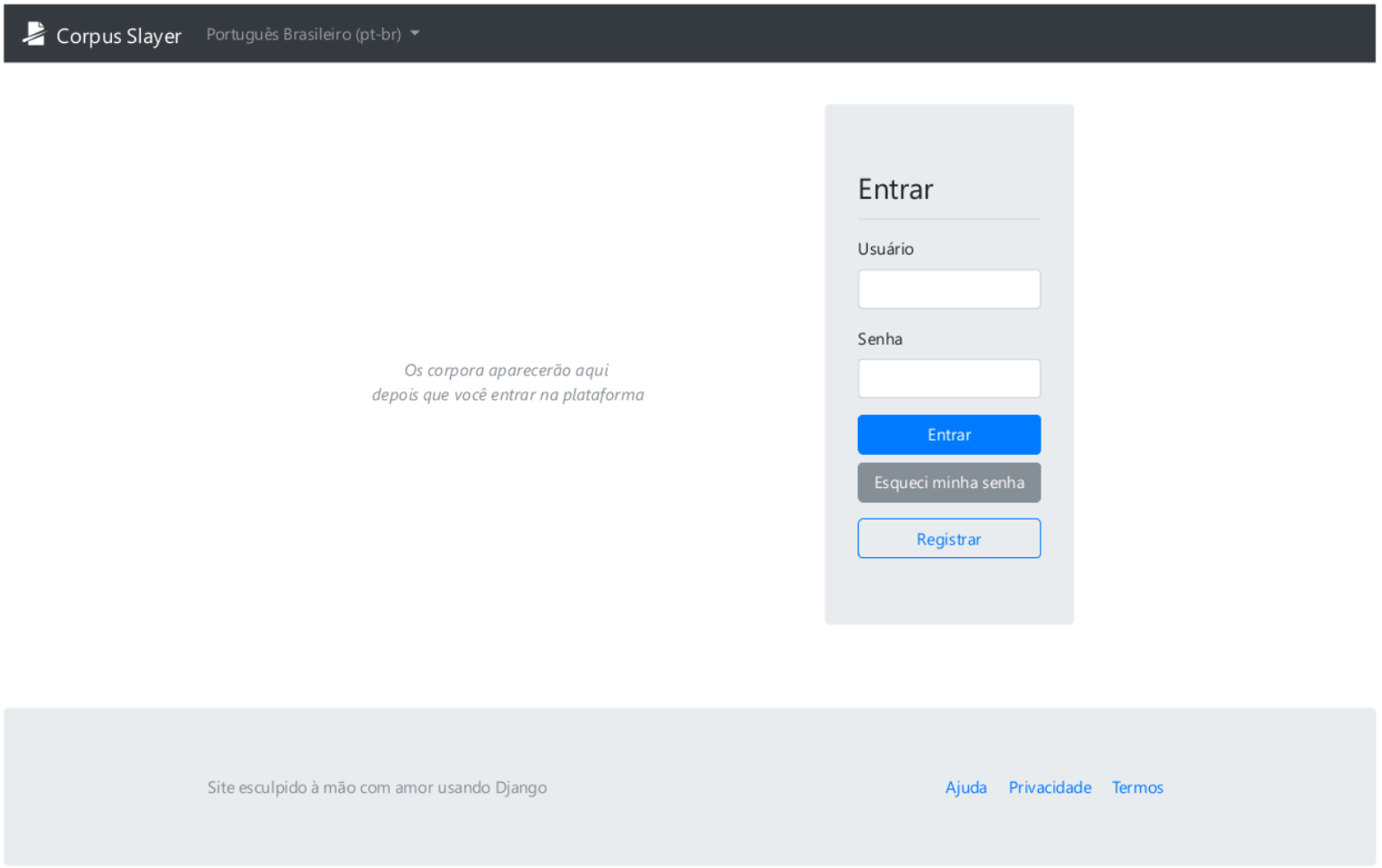
Etiquetador	Precisão
MXPOST	89, 66%
Brill Tagger	88, 76%
Tree Tagger	88, 47%
YAS-Tagger	76, 93%
Unitex/GramLab	60, 76%

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

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

S

O sistema desenvolvido



Tela inicial do sistema desenvolvido

Fonte: O autor

S

O sistema desenvolvido

Seletor de idiomas do sistema desenvolvido
Esquerda: *desktop*
Direita: *mobile*

Fonte: O autor

S

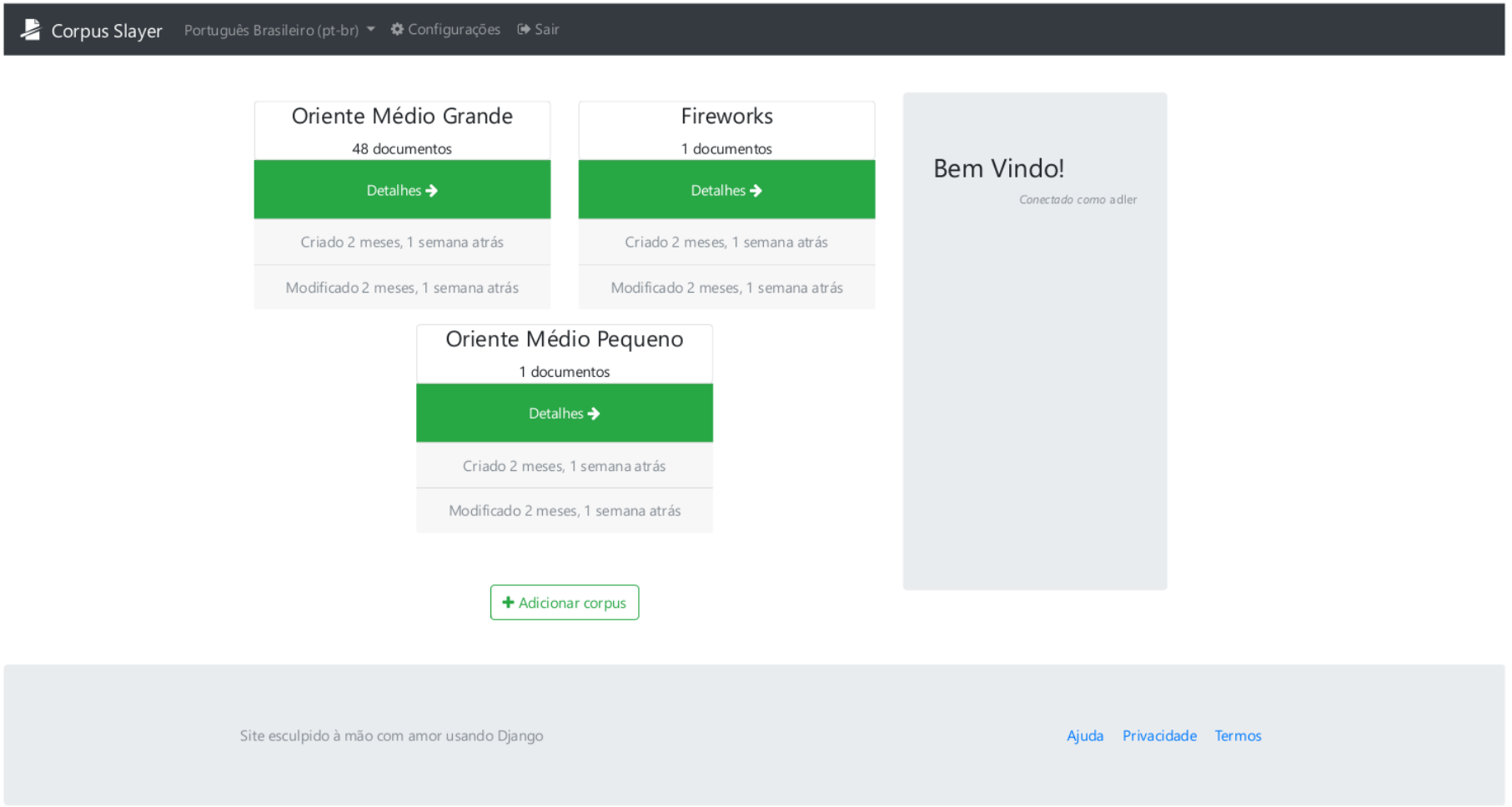
O sistema desenvolvido

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

Fonte: O autor

S

O sistema desenvolvido

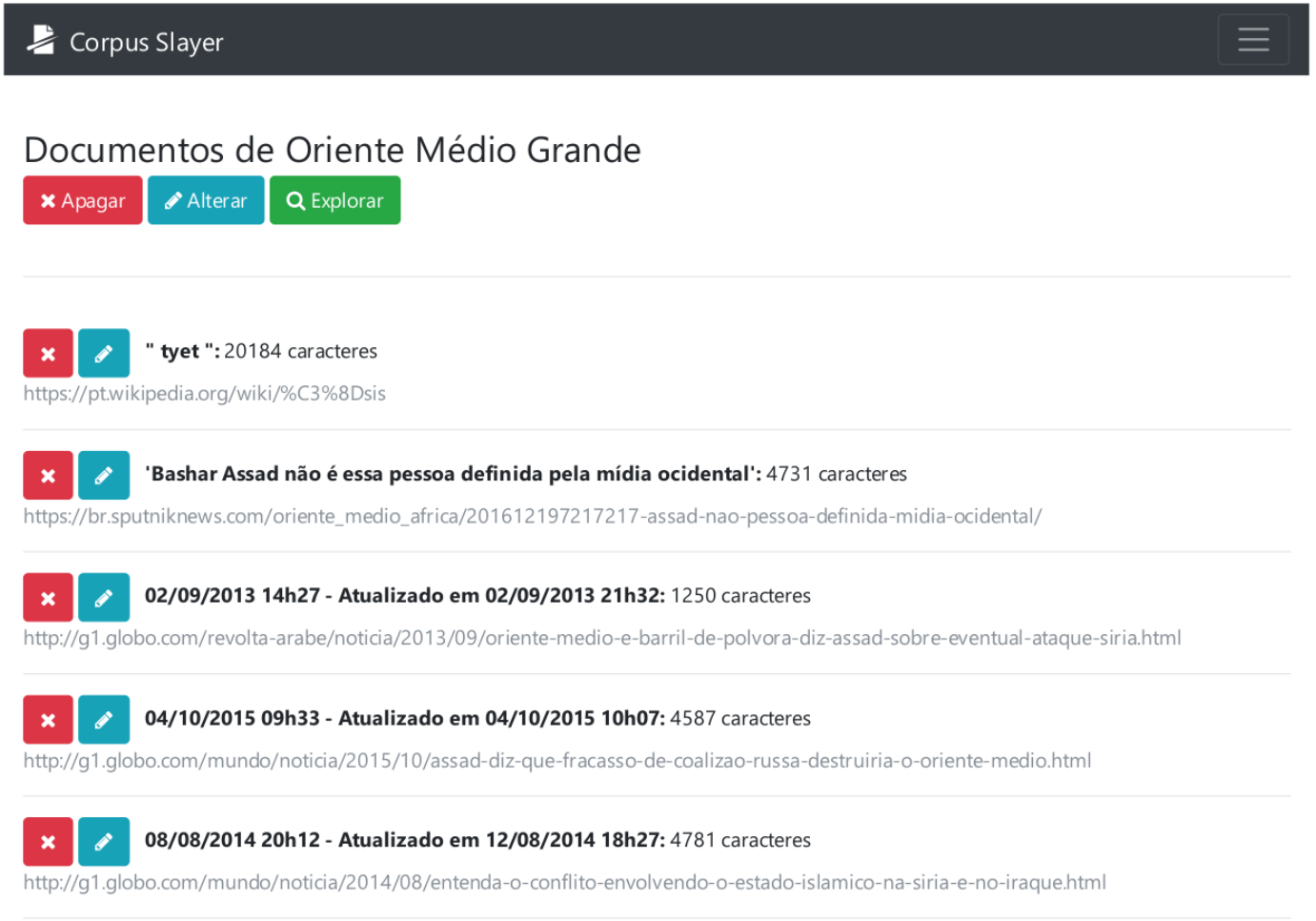


Tela inicial do sistema desenvolvido, após autenticação, exibindo a lista de corpora

Fonte: O autor

S

O sistema desenvolvido



Lista de documentos dum corpus

Fonte: O autor

S

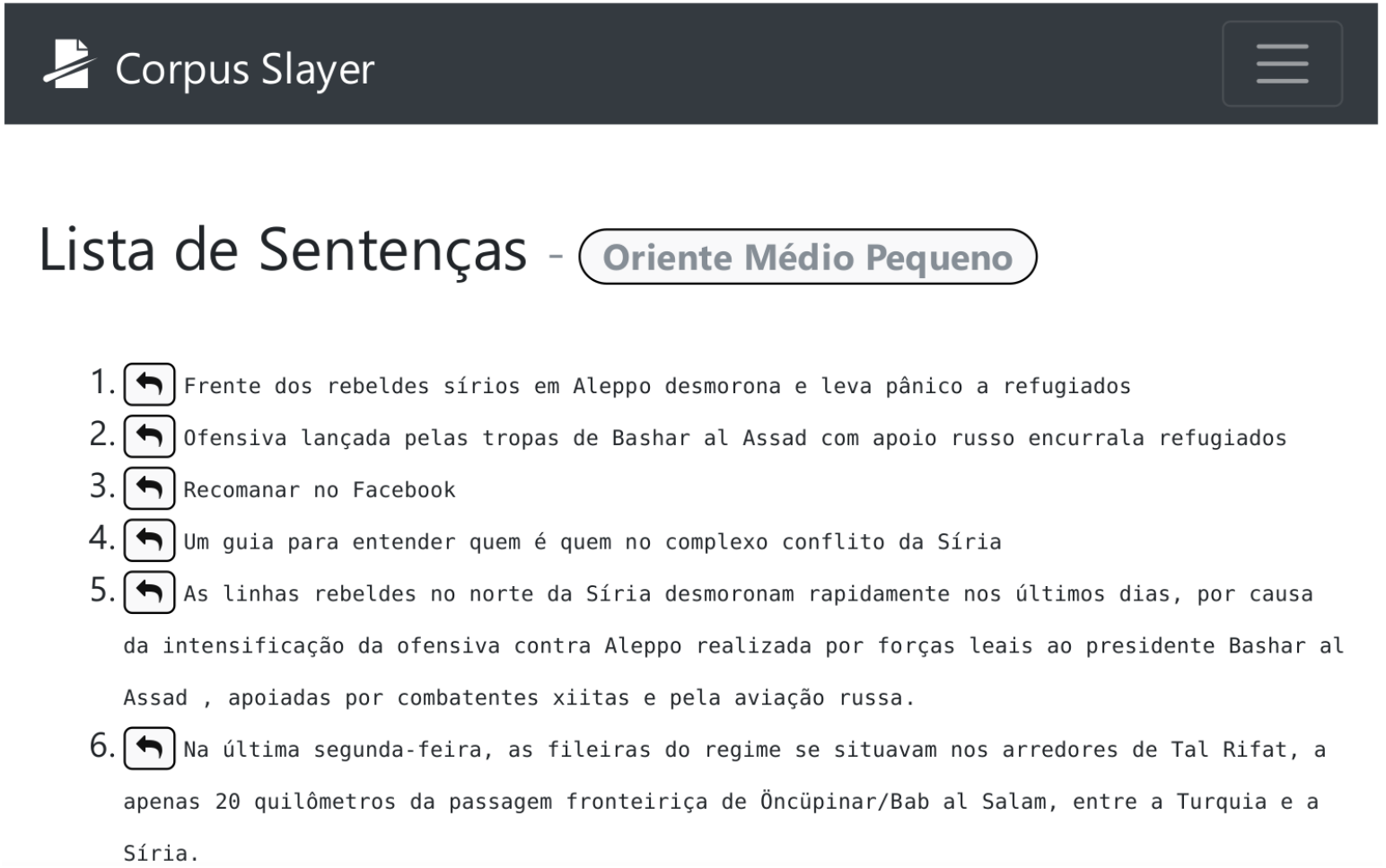
O sistema desenvolvido

Opções de análises para um corpus

Fonte: O autor

SS

Extrator de sentenças e lista de palavras

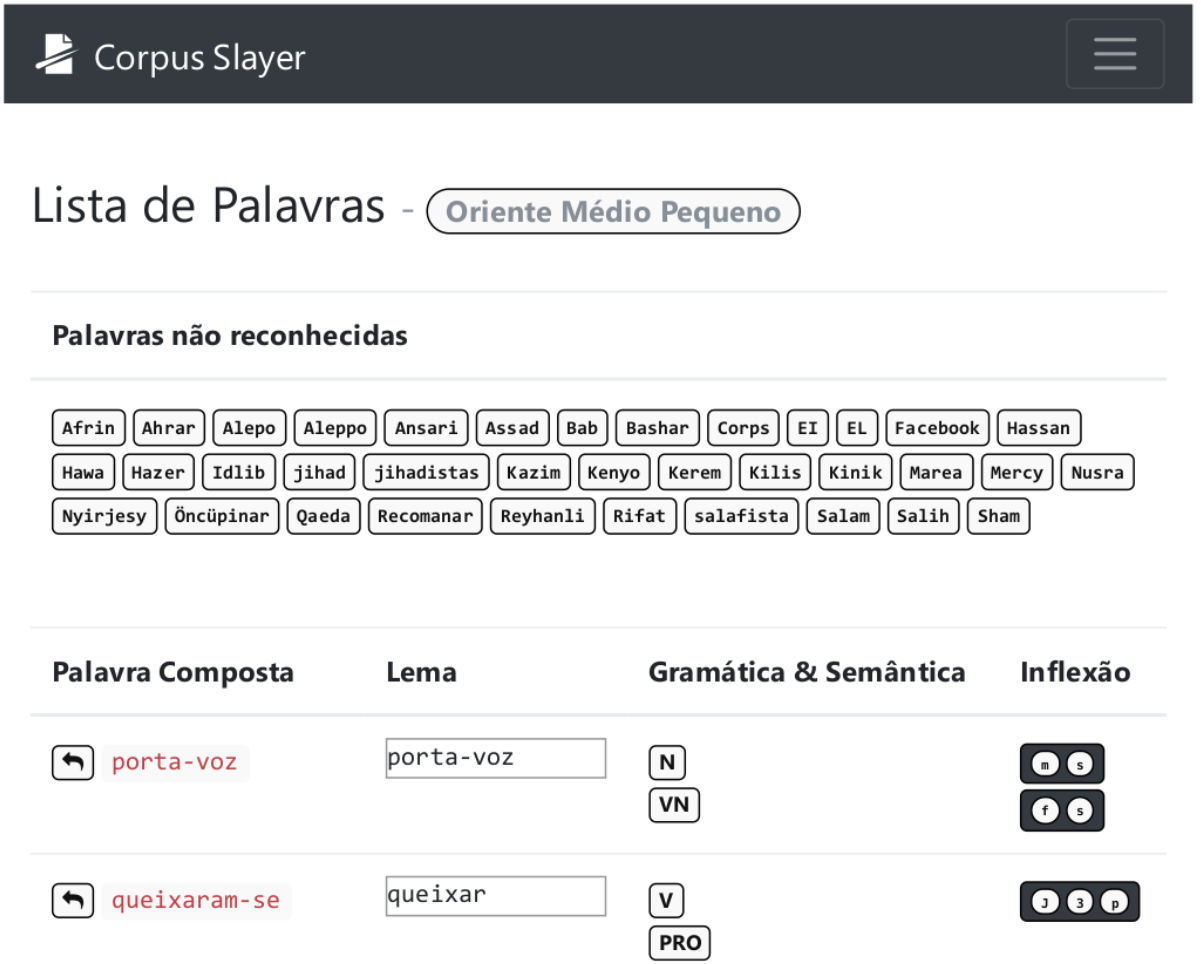


Lista de sentenças dum corpus

Fonte: O autor

SS

Extrator de sentenças e lista de palavras

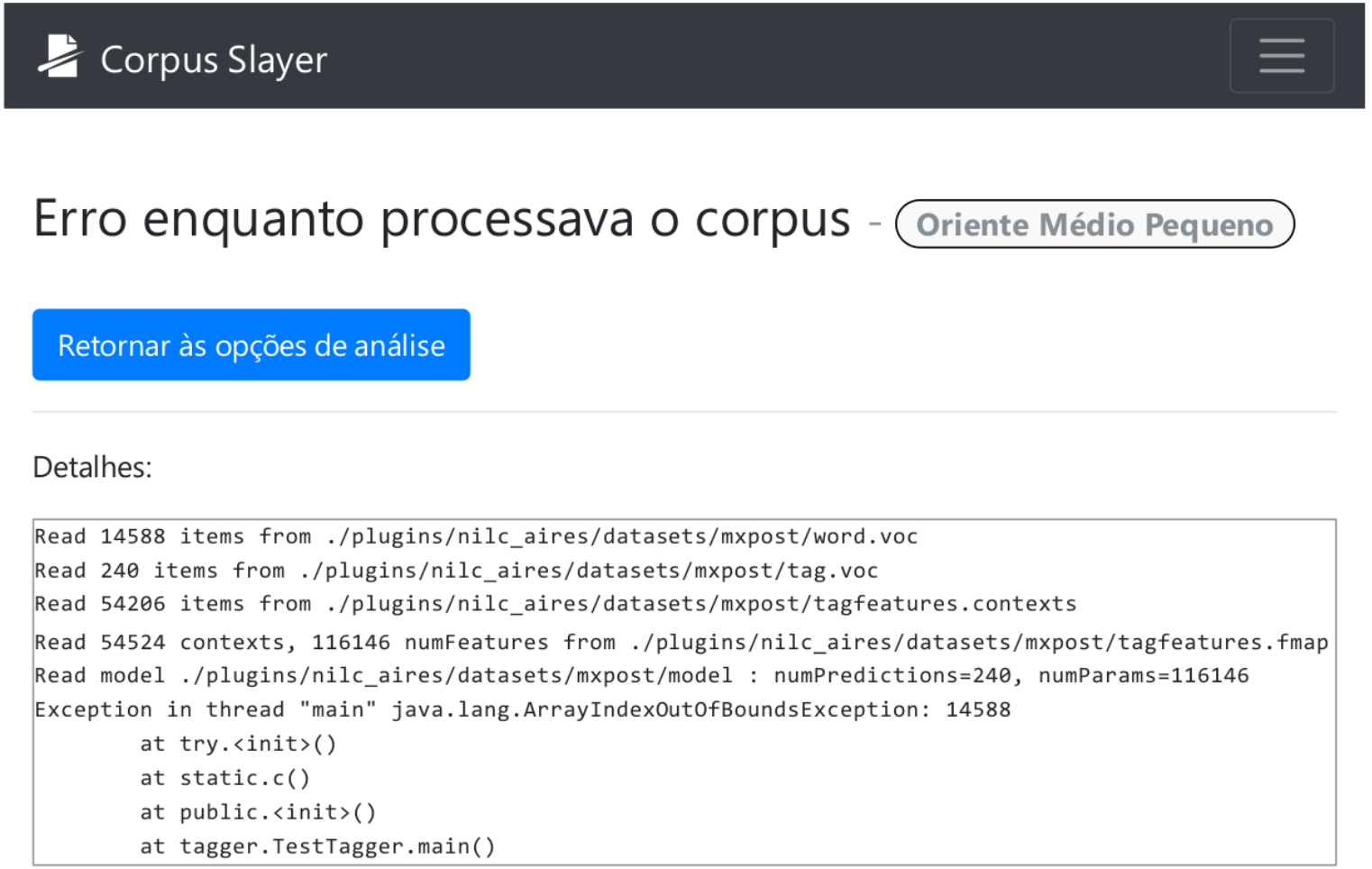


Lista de palavras dum corpus

Fonte: O autor

SS

Etiquetador



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*

SS

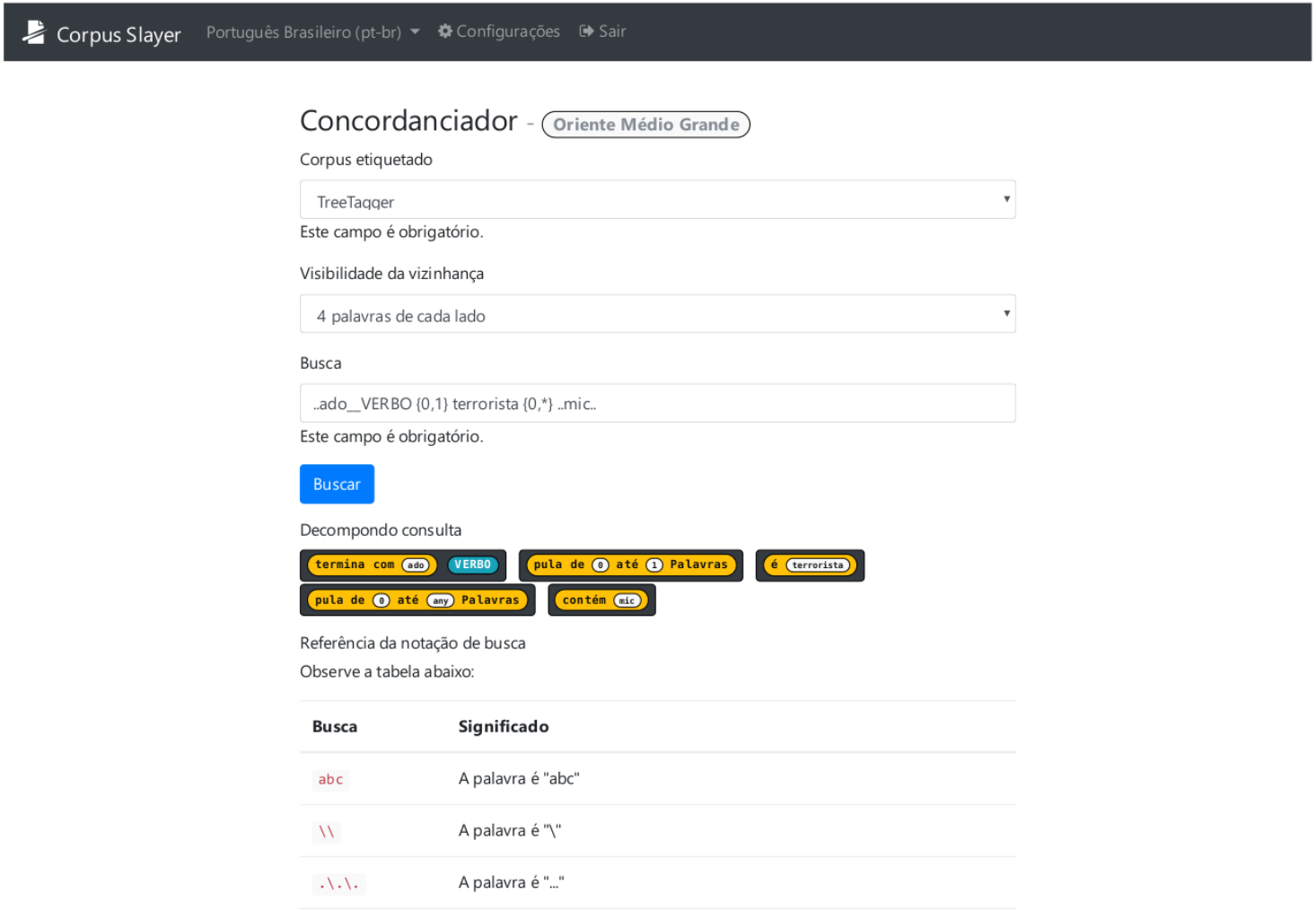
Etiquetador

Corpus processado pelo *Tree Tagger*

Fonte: O autor

SS

Concordanciador



Tela de busca do concordanciador

Fonte: O autor

SS

Concordanciador

Tela de resultados do concordanciador

Fonte: O autor

SS

A “seta para trás”

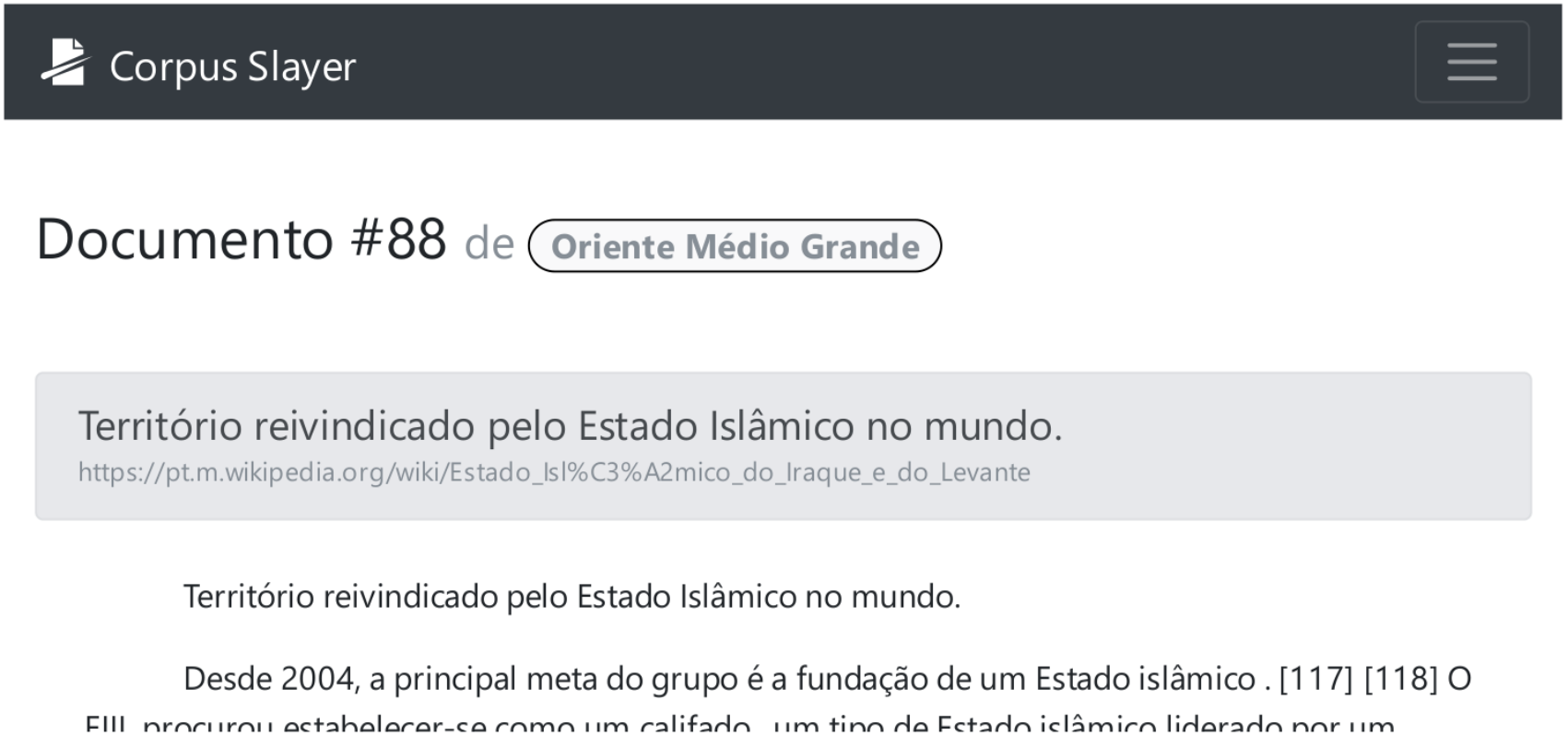
A “seta para trás” presente na tela de resultados do concordanciador, na do etiquetador e nas listas de sentenças e lista de palavras, em destaque

Fonte: O autor

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

SS

A “seta para trás”



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

Fonte: O autor

SS

A “seta para trás”

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

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

Fonte: O autor

C

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.

S

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.

Dúvidas?