Criação de um Ambiente para o Processamento de Córpus de Português Histórico

Arnaldo Candido Junior

Orientadora:
Sandra Maria Aluísio

CTDIA - Salvador - 25 de Outubro de 2008

Roteiro

- Contexto do Trabalho de Mestrado
  - O projeto DHPB
- Objetivos
- Desenvolvimento do trabalho
  - Córpus e Glossários
  - Ferramentas
  - Ambiente
- Conclusão

Cenário de usos de Córpus

- Lingüística (fonética, lexicografia, sintaxe, semântica, etc)
- Processamento de Língua Natural
- Construção de ferramentas e recursos (engenharia da linguagem)
- Ensino de idiomas
- Sociologia e História

O projeto DHPB

- Dicionário Histórico de Português do Brasil
  - O primeiro do gênero
  - Português brasileiro começou a divergir do Português europeu (cultura, fauna, flora)
  - Período pré-imprensa do Brasil (1500-1808)
- Baseado em um grande córpus de textos históricos
  - Autores brasileiros ou portugueses que viveram no Brasil por um longo período

Objetivos do Trabalho

- Construção de um ambiente para o processamento de córpus históricos
- Aplicado a um projeto maior (DHPB) para a construção de um dicionário de português histórico
- Metodologias, recursos e ferramentas

Carta do padre Manuel da Nóbrega ao padre Simão Rodrigues (1549)

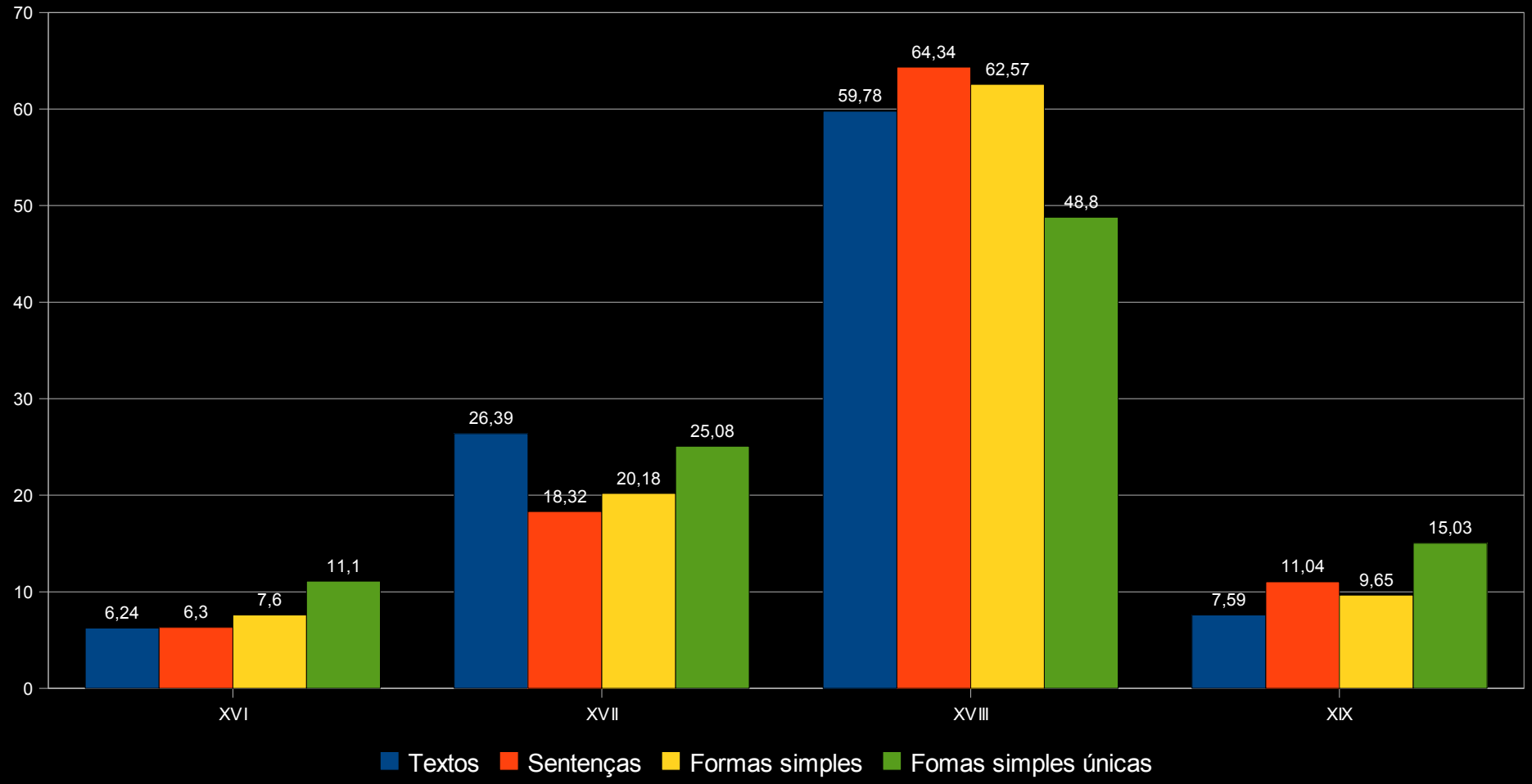
convertidos, onde estaremos Vicente Rodriguez e eu, e hum soldado[19] que se meteo comnosco para nos servir, e está agora em Exercicios, de que eu estou muy contente. Faremos nossa igreja, onde insinemos os nossos novos christãos, e aos domingos e festas visitarey a Cidade e pregarey.

O Padre Antonio Pirez e o P.e Navarro estaram em outras Aldeas longe, onde já lhes fazem casas. E portanto hé necessario V. R. mandar officiaes, e am-de vir já com a paga, porque cá diz ho Governador que, ainda que venha Alvará de S. A. para nos dar o necessario, que nom o averá hi para isto. Os officiaes que cá estão tem muito que fazer, e que o nom tenhão, estão com grande saudade do Reyno, porque deixão lá suas molheres e filhos, e nom aceitaram a nossa obra depois que cumprirem com S. A., e tambem ho trabalho que tem com as viandas e o mais os tira disso. Portanto me parece que avião de vir de lá, e, se possivel

Dados do Córpus

- 2.5 mil textos
- 7.5 milhões de palavras
- 369 mil palavras únicas
- 82 MB (UTF-16LE)

Textos por século (%)



Desafios encontradas no projeto DHPB

- Metadados (uso do padrão TEI)
- Ausência de hifenização (tratamento manual)
- Símbolos tipográficos incomuns (uso do Unicode)
- Junções de palavras como “éamor” (criação de manual de um glossário)
- Abreviaturas (glossário com informações do dicionário FLEXOR)
- Variações de grafia (detecção automática – dicionários)
- Problemas similares aos levantados por Rydberg-Cox (2003)

Símbolos tipográficos incomuns

Acentos combinados

Símbolo	Unicode	Freqüência	Exemplo
◌̂	0302	2	quarŷ
◌̃	0303	24.568	coñercio
◌̄	0304	596	cacaō
◌̈	0308	48	muÿ
◌̉	0309	1.804	sỏmente
◌̓	0313	371	tinhaỏ
◌́	0301	2	quaeś
◌̆	0306	2	apanhĕ

Símbolos em geral

Símbolo	Unicode	Freqüência	Exemplo
Æ	00C6	41	Æthyopia
æ	00E6	1.378	græti
œ	0153	116	Cœteris
§	00A7	1.131	§ (parágrafo)
ⅎ	2132	4	ⅎixit
ſ	017F	928	Deſcobrio (*)
ƒ	0192	149.909	ƒeito (*)
ɐ	0250	4	passadɐ
&	0026	20.649	&c. (etc.)
@	0040	192	@nrique

* Freqüência sujeita a revisão

Abreviaturas, junções, variantes

Exemplo	**Abreviatura**	**Junção**	**Variante**
Sarg.^{José} (Sargento José)	X	X	
abaſe (a base)		X	X
Supp^{te} (supostamente)	X		X
héalagadacomm^{tos} (é alagada com muitos)	X	X	X

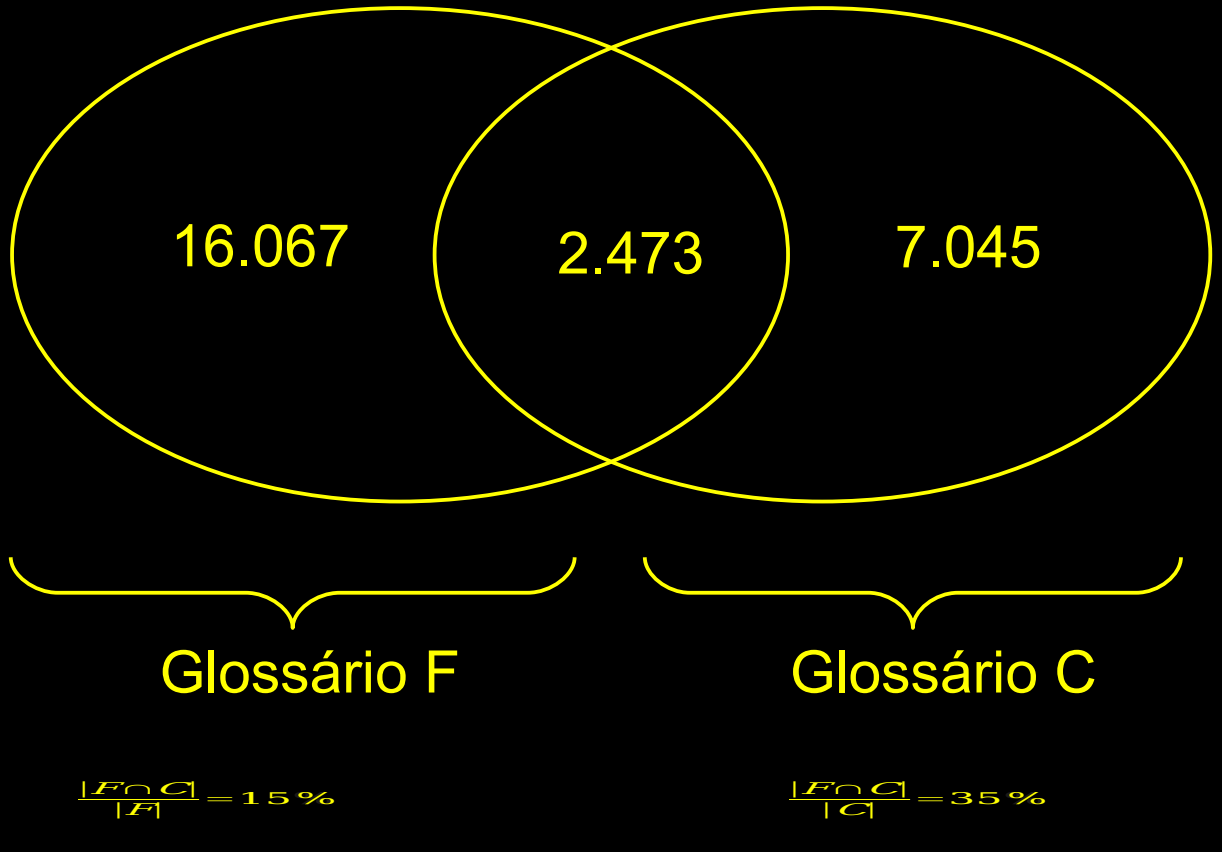
Abreviaturas

- Expansões de Bº: bairro, Bartolomeu, bastardo, beco, bento, Bernardo, etc.
- Abreviaturas de janeiro: jan., jan.ro, janro, janr.º, jan.º, etc.
- Glossário F: criado manualmente a partir de Flexor (1991)
- Glossário C: extraído do córpus através de heurísticas

Heurísticas

- Presença de sobrescrito: ant.º, cid.ª, p.ª (61%)
- Ponto interno sucedido por até 4 símbolos: cid.e, embg.e, ex.mo (24%)
- Palavras terminadas por algumas consoantes: cap, reg, liv, v (15%)

Abreviaturas



$\frac{|F \cap C|}{|F|} = 15\%$

$\frac{|F \cap C|}{|C|} = 35\%$

Variantes de grafia

- Inexistência de um sistema ortográfico unificado em textos históricos
- Exemplo: villa, vyla, vjlla, vylla, vjla
- Abordagem 1: regras de transformação (Giusti et al., 2007)
- Abordagem 2: Distância de edição (Philologic + Agrep)

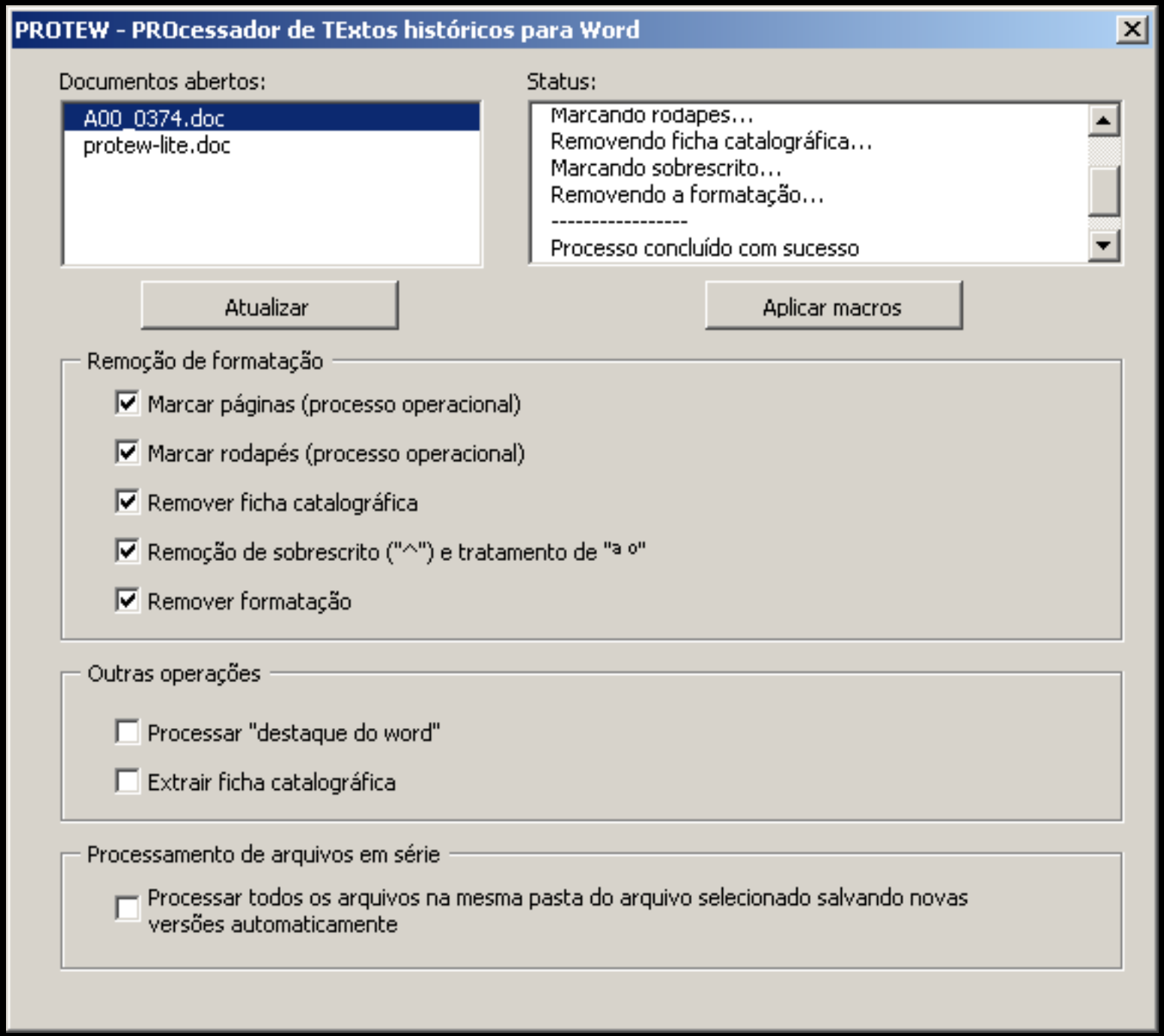
Regras de transformação

- Exemplo: (e[ao] e ei)
  - "e[ao]" cobre “aldea”, “meo”, “cheas”, etc
  - "e" subcadeia a ser substituída (aldea, meo, cheas, etc)
  - "ei" cadeia de substituição (resultado em aldeia, meio, cheias, etc)
- 43 regras de transformação, 18.082 agrupamentos e 41.170 variantes

Ferramentas

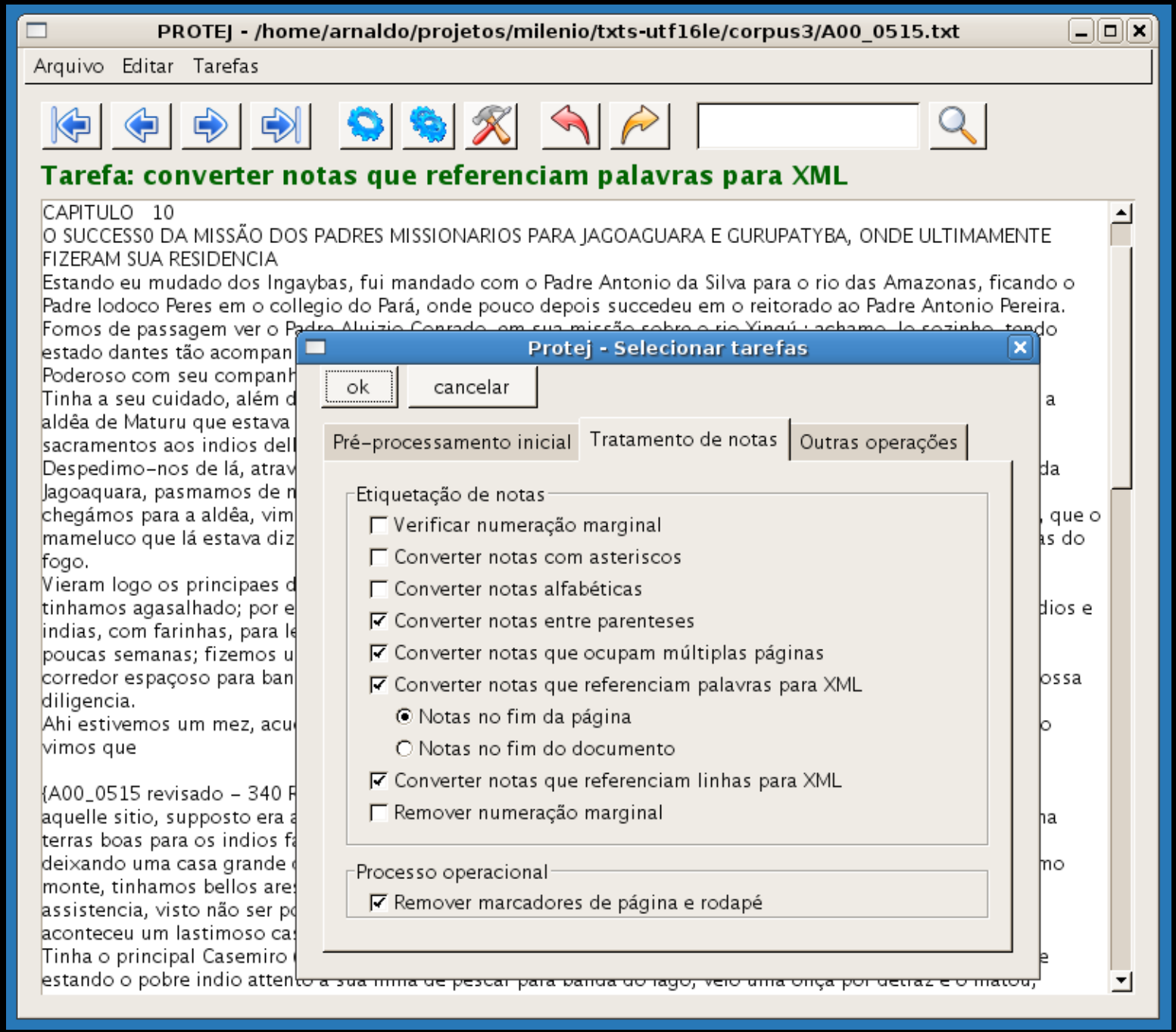
- Protew: pré-processamento do córpus *
- Protej: pré-processamento do córpus e dos glossários *
- Siaconf: detecção de variantes de grafia
- Unitex: acesso a córpus
- Philologic: acesso a córpus
- Procorph: redação de verbetes *
- * nilc.icmc.usp.br/nilc/projects/procorph/

Protew-lite

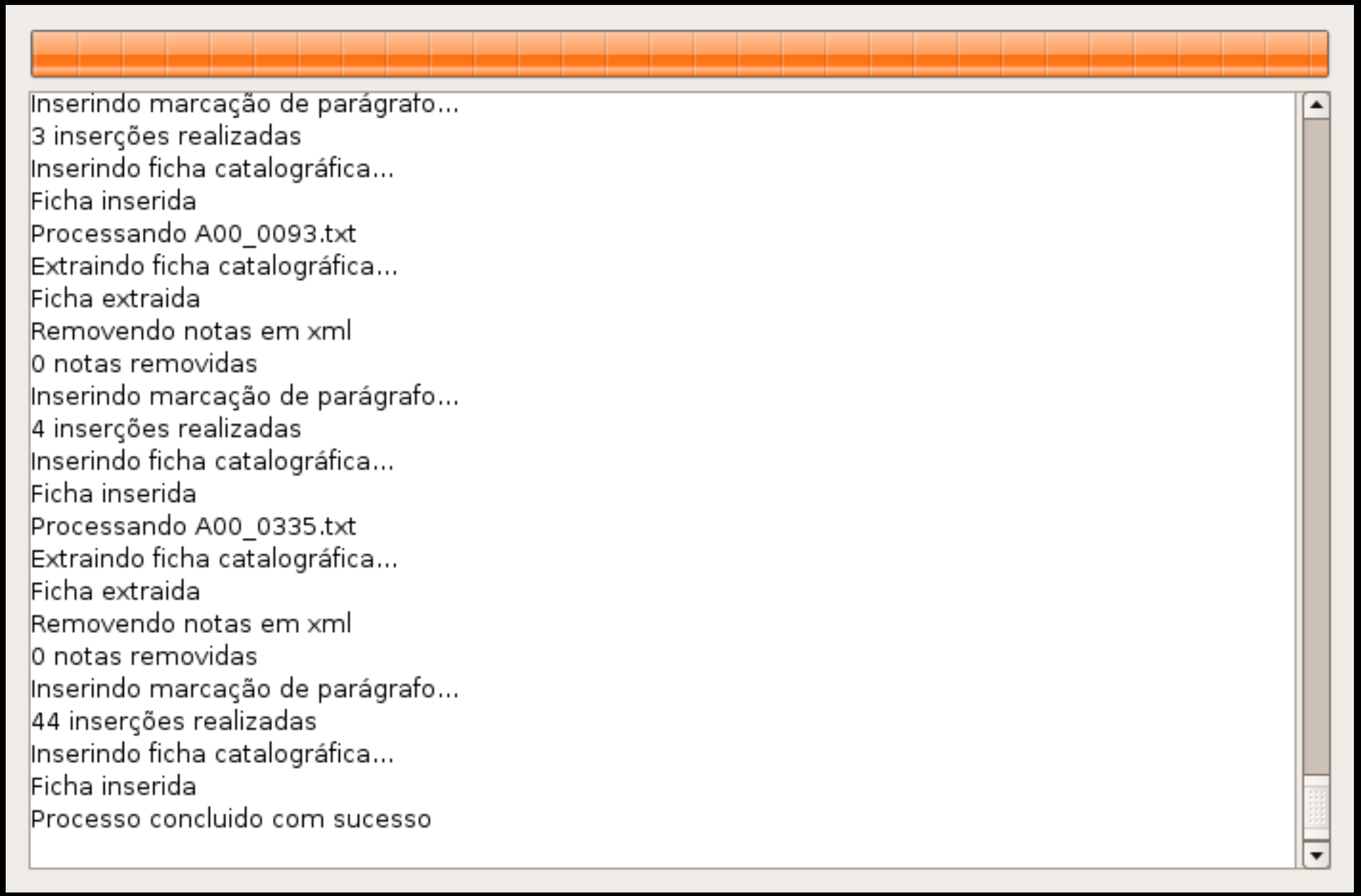

PROTEW - PROcessador de TExtos históricos para Word
Documentos abertos:
A00_0374.doc
protew-lite.doc
Status:
Marcando rodapes...
Removendo ficha catalográfica...
Marcando sobrescrito...
Removendo a formatação...

Processo concluído com sucesso
Atualizar
Aplicar macros
Remoção de formatação
Marcar páginas (processo operacional)
Marcar rodapés (processo operacional)
Remover ficha catalográfica
Remoção de sobrescrito ("^") e tratamento de "a o"
Remover formatação
Outras operações
Processar "destaque do word"
Extrair ficha catalográfica
Processamento de arquivos em série
Processar todos os arquivos na mesma pasta do arquivo selecionado salvando novas versões automaticamente

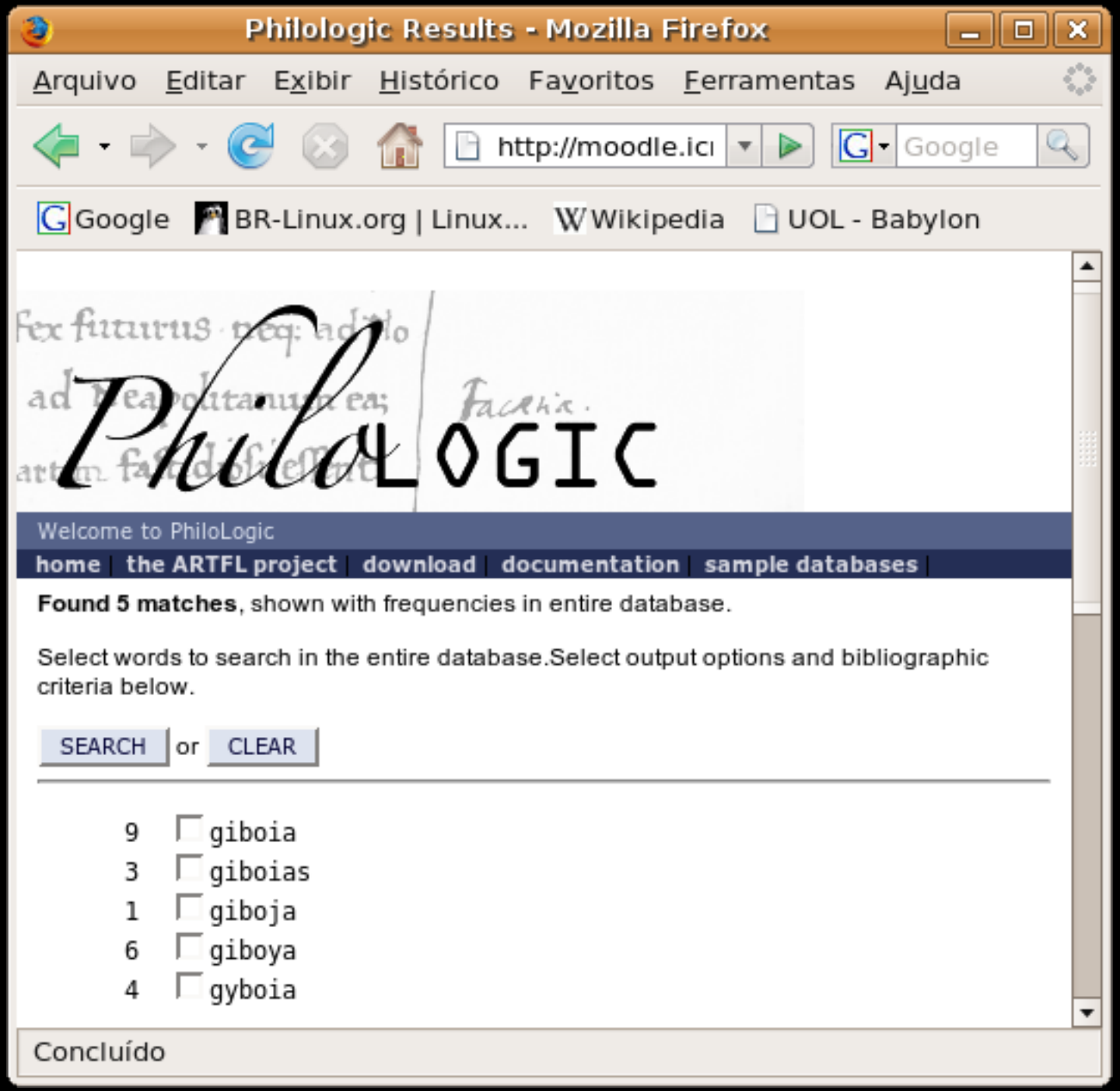
Protej


PROTEJ - /home/arnaldo/projetos/milenio/txts-utf16le/corpus3/A00_0515.txt
Arquivo Editar Tarefas
Tarefa: converter notas que referenciam palavras para XML
CAPITULO 10
O SUCCESSO DA MISSÃO DOS PADRES MISSIONARIOS PARA JAGOAGUARA E GURUPATYBA, ONDE ULTIMAMENTE FIZERAM SUA RESIDENCIA
Estando eu mudado dos Ingaybas, fui mandado com o Padre Antonio da Silva para o rio das Amazonas, ficando o Padre Iodoco Peres em o collegio do Pará, onde pouco depois succedeu em o reitorado ao Padre Antonio Pereira.
Protej - Selecionar tarefas
ok
cancelar
Pré-processamento inicial
Tratamento de notas
Outras operações
Etiquetação de notas
Verificar numeração marginal
Converter notas com asteriscos
Converter notas alfabéticas
Converter notas entre parenteses
Converter notas que ocupam múltiplas páginas
Converter notas que referenciam palavras para XML
Notas no fim da página
Notas no fim do documento
Converter notas que referenciam linhas para XML
Remover numeração marginal
Processo operacional
Remover marcadores de página e rodapé

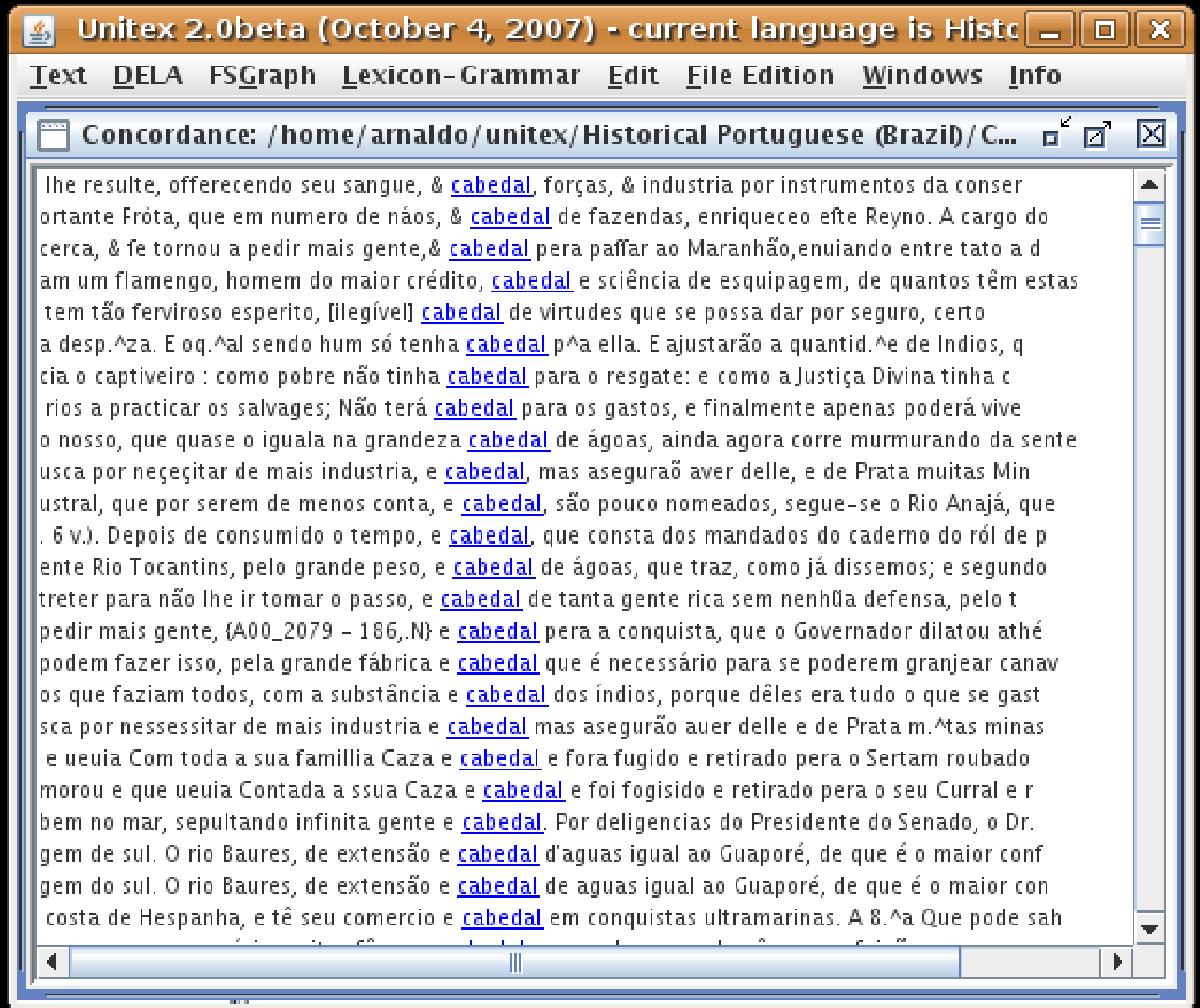
Protej – Córpus Philologic



Philologic

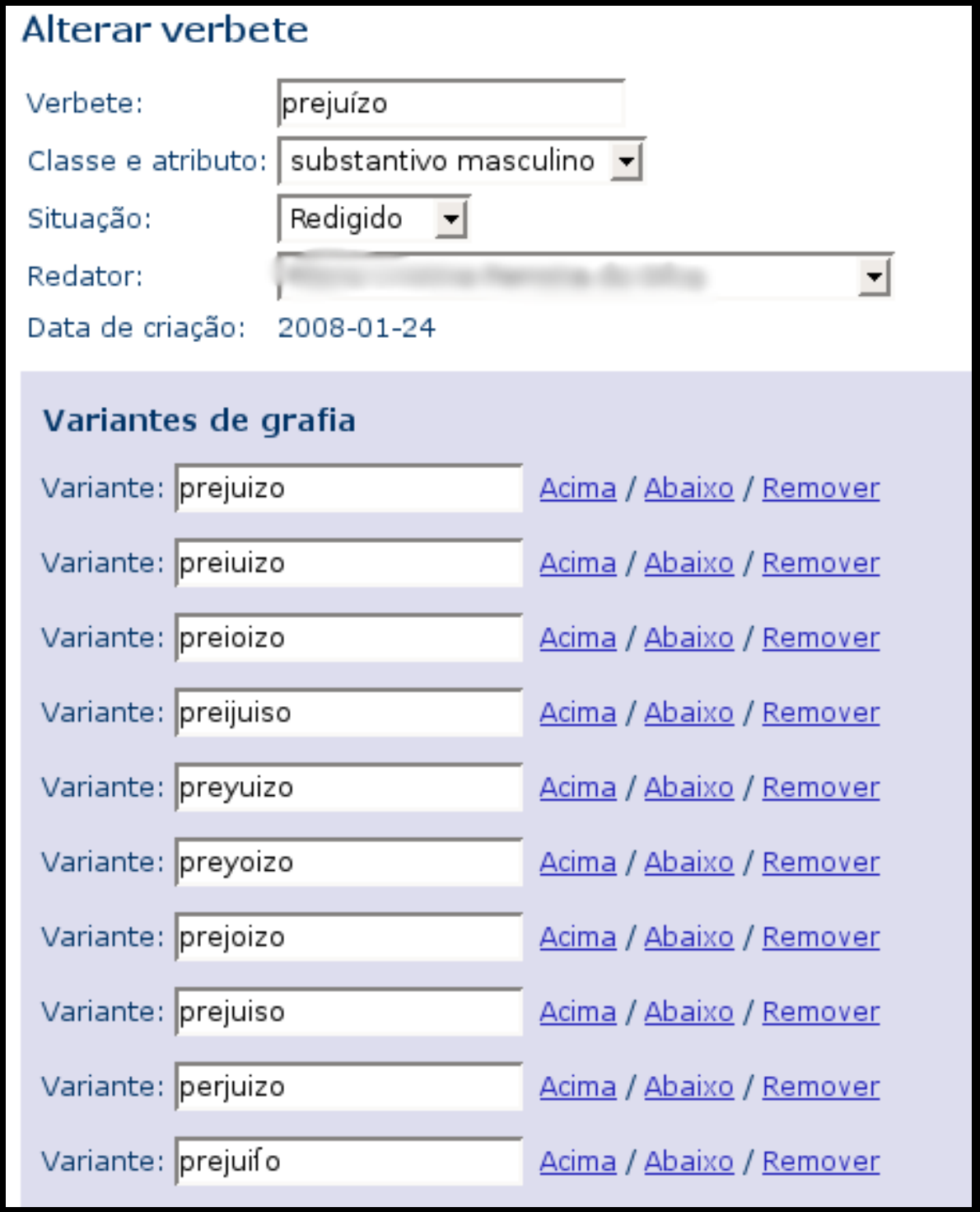

Philologic Results - Mozilla Firefox
Arquivo Editar Exibir Histórico Favoritos Ferramentas Ajuda
http://moodle.ici
Google
Google BR-Linux.org | Linux... Wikipedia UOL - Babylon
PhiloLOGIC
Welcome to PhiloLogic
home | the ARTFL project | download | documentation | sample databases |
Found 5 matches, shown with frequencies in entire database.
Select words to search in the entire database.Select output options and bibliographic criteria below.
SEARCH or CLEAR
9 giboia
3 giboias
1 giboja
6 giboya
4 gyboia
Concluído

Unitex
http://www-igm.univ-mlv.fr/~unitex/



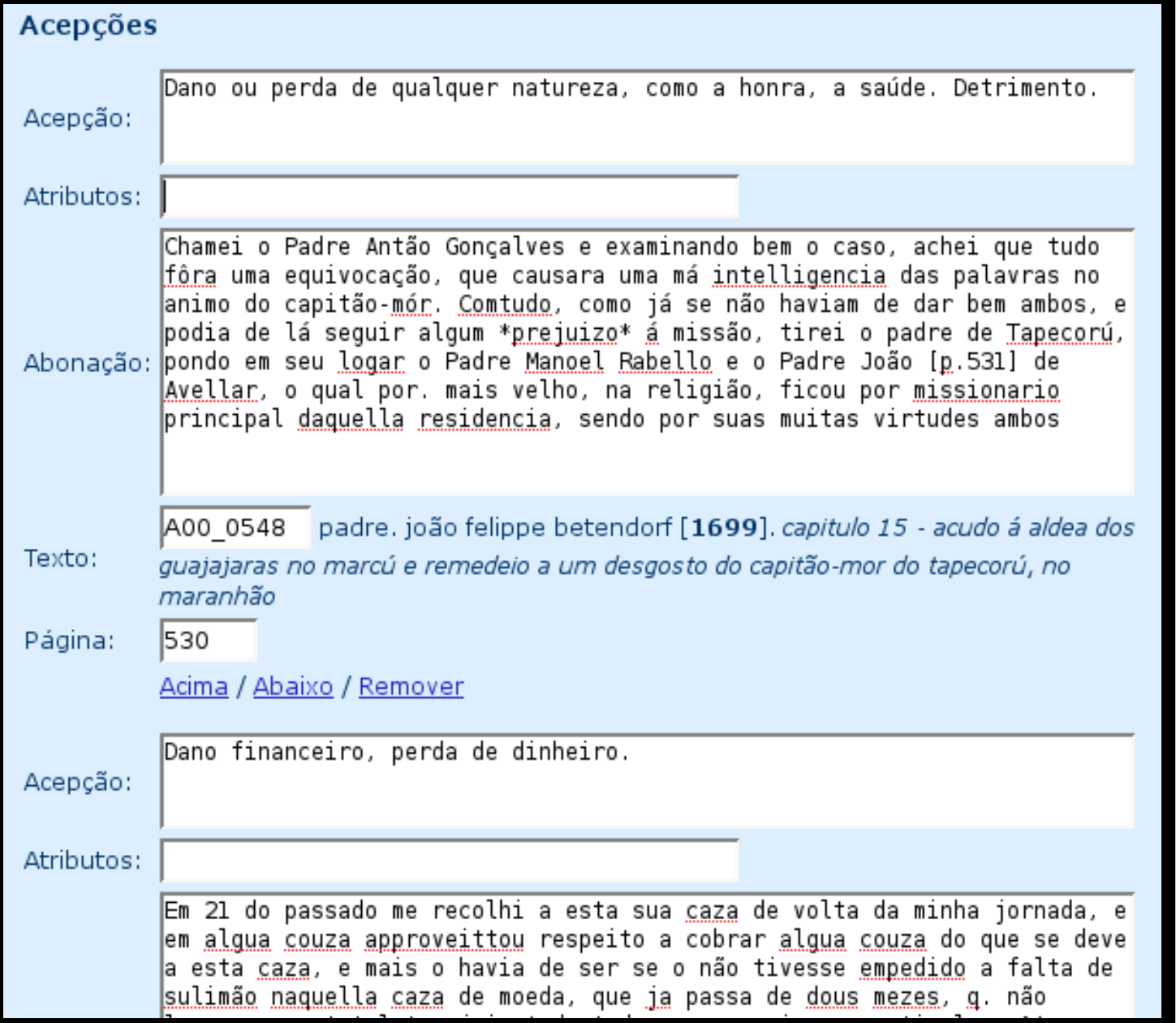
Procorph

- Dados gerais
- Variantes



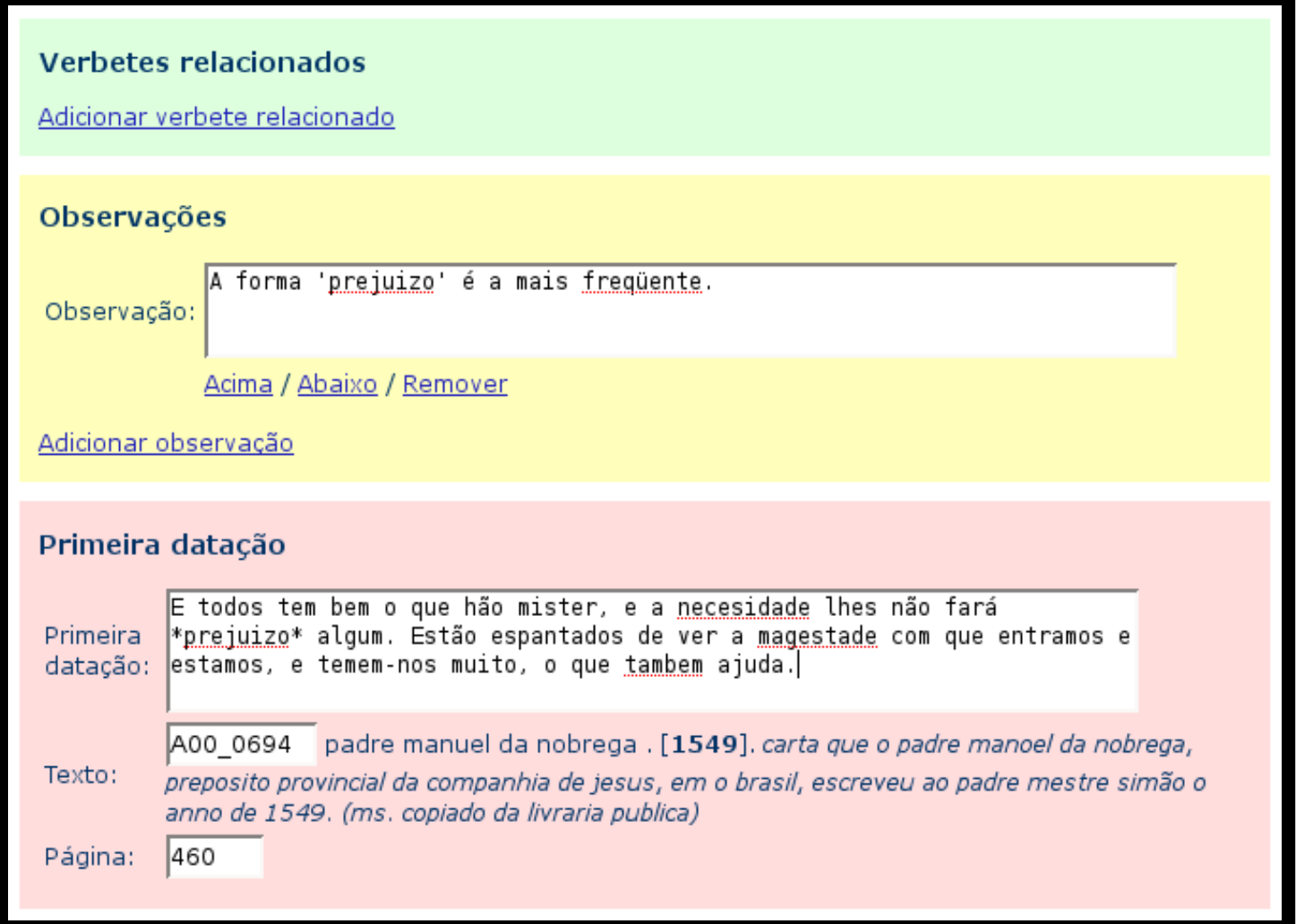
Procorph

- Abonações
- Acepções
- Referências



Procorph

- Relacionados
- Observações
- Primeira datação



Procorph

- Visualização (açucar)

açúcar: substantivo feminino.

Variantes: assúcar, asúcar

1. Substância doce fabricada industrialmente, extraída, em geral, da cana de açúcar.

> Porque quem vos mostrara, há hoje trezentos anos, uma cana de que se faz o **açúcar** e vos dissera que daquela cana se havia de formar com a indústria humana um pão de **açúcar** tão formoso como hoje o vemos, tê-lo-ieis por cousa ridiculosa; e, pelo conseguinte, se vos fôsse mostrado um pedaço de pano velho de linho e vos afirmassem que daquele pano se havia de fazer o papel em que escrevemos, quem duvida que o terieis por zombaria? **ambrósio fernandes brandão [1618].** ***diálogo primeiro*****, p. 1** .

Primeira datação: A renda, que El-Rei cá tem nesta Baía, é esta, scilicet: as miunças que rendem cento e vinte mil réis em que andam arrendadas; o peixe e mandioca e algodão andam em cento e trinta mil réis; pagos em ordenado, que é um terço menos, pode valer em dinheiro oitenta mil réis; o **açúcar** do Engenho anda em cento e cincoenta cruzados. **p. manuel da nóbrega [1558].** ***carta do p. manuel da nóbrega ao p. miguel de torres, baía 8 de maio 1558*****, p. 26**.

Concordâncias

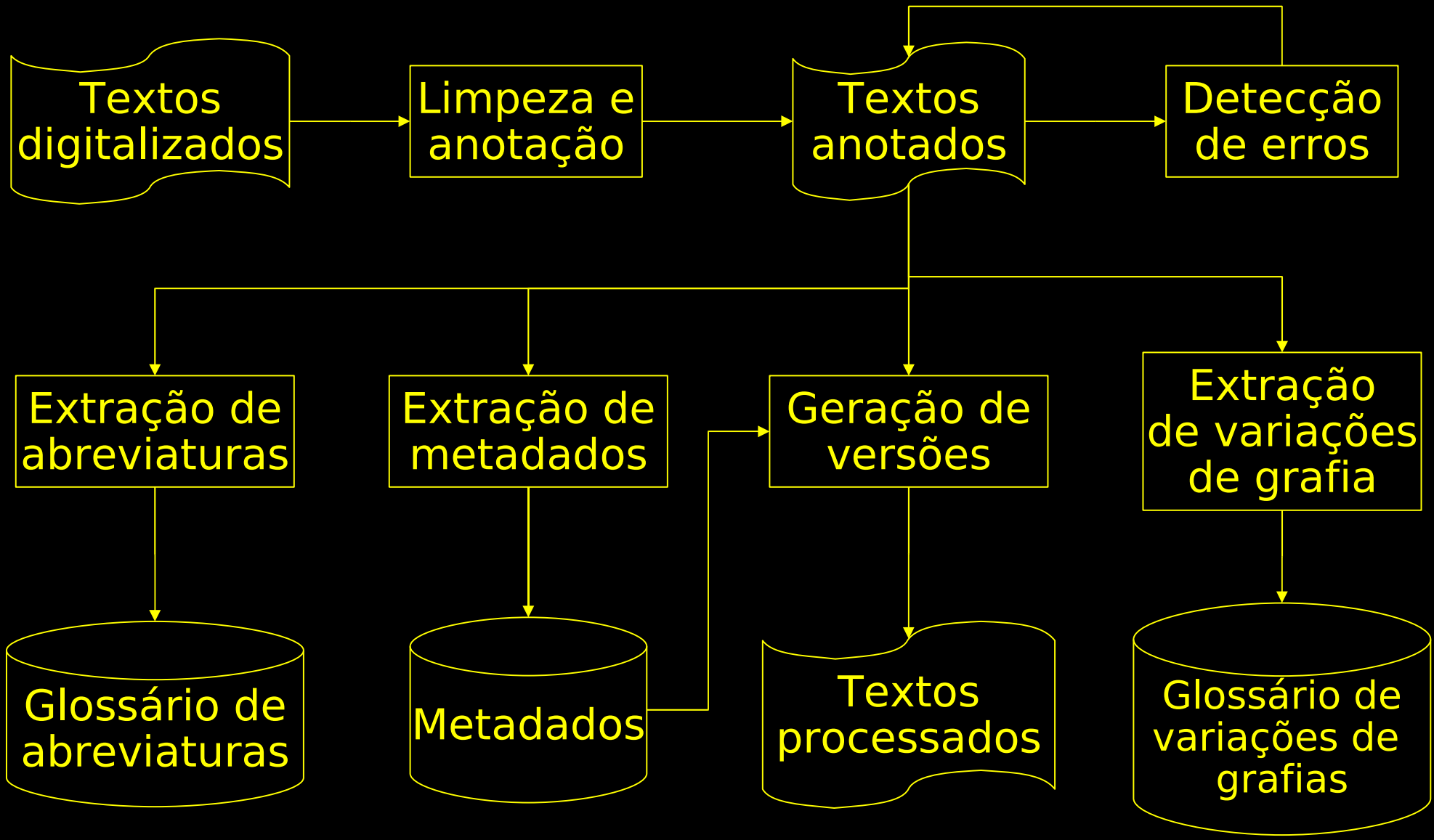
Busca por concordâncias

Buscar por: cabedal [Buscar]

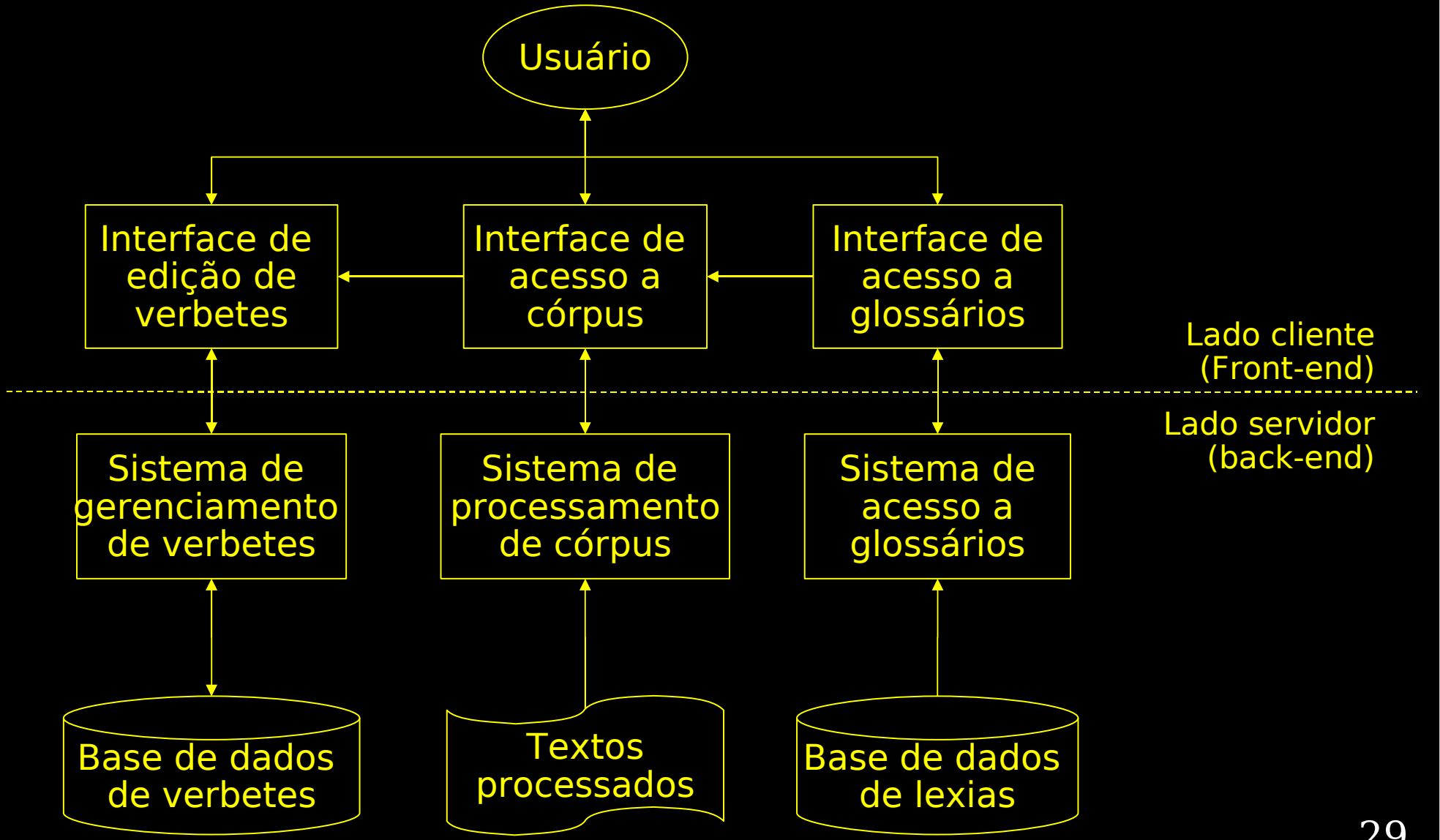
<< Início < Voltar 201 a 300 de 378 Avançar >

das quais houve intensidade Tem êste rio bastante **cabedal** de águas cristalinas e saborosas Um sucesso dêste
êste lugar da cachoeira do Ribeirão um rio assaz **cabedal** de boa água com uma laje de mais de passos
nteza grande o que concorre para representar mais **cabedal** do que o próprio Guaporé nesta foz Também nos falt
rdade a huma multidaõ de Indios q̃ fazia o grande **Cabedal** dos Paulistas e com q̃ de necessidade em hum dia
ras visto q̃ os Índios as naõ lavraraõ nem tinhaõ **Cabedal** p ^a isso E q̃ p ^a o dito effeito
efeitorio o qual serve denfermaria estando o dito **cabedal** enterrado em panellas de cobre e por cima pedra mi
de oiro Enquanto os Taberneiros ajuntavam Imenso **cabedal** em poucos anos CARTAS CHILENAS Sem terem nas Taber
to Tiraste de gastar em frias festas CARTA Imenso **cabedal** que o bom Senado Devia consumir em cousas santas S
De muito mais abono e a quem devia Um grosso **cabedal** o Régio Erário Mal acaba Marquésio o seu triênio O
fe Mostrar um grande zelo nas cobranças Do imenso **cabedal** que todo o povo Aos Cofres do Monarca está devendo
Exclama o bom Ribério que não pode Pois todo o **cabedal** que tem cobrado Ou está nas demandas consumido Ou
ella importante e larga fronteira como pelo maior **cabedal** de agoas destes grandes rios que facilitam o chega
a do Pará se faz para VillaBella tanto pelo maior **cabedal** de agoas dos Amazonas Madeira Mamoré e Guaporé em
que os primeiros descobridores vendo o seu maior **cabedal** de aguas tomaram pelo principal rio conflue com a
oré pela mesma margem de norte rio de não pequeno **cabedal** daguas Do Cautarios são legoas de navegação a rumo
sua margem de sul O rio Baures de extensão e **cabedal** daguas igual ao Guaporé de que é o maior confluent
sendo o Mamoré rio de grande largura e de maior **cabedal** daguas elle traz as suas origens da latitude de gr
p^a ajuda das expensas desta expediçaõ fiz do meo **Cabedal** dezejozo de que exforçandose os demais Vassalos de
metendose nos mesmos donde nascerão e com todo o **cabedal** de agoas que abraçarão se dirigem as ditas barras
ados braços que tomão diversos nomes e com todo o **cabedal** de agoas desembocão em suas Barras deixando de cam
soldados Ama a gente assisada a honra a vida o **cabedal** tãopouco que ponha uma acção destas nas mãos dum p
os meus amores Eu ó cega não tenho um grosso **cabedal** dos pais herdado não o recebi no emprego nem tenho
o seacha Excrito choice sic decabedaL sic corr de **cabedaL** corr choice pera poder acudir ao sus tento da infa
CÂMARA esta Cidade athe o tempo prezente enam tem **cabedal** para poder alugar Cazas epagar pelas razoens asima
the aqui tem Sucedido havendo pessoas demais poço **Cabedal** que há doze annoz selhetem tomado suas Cazas choic
onforme os mesmoz Capitaens ordenarem etiverem de **Cabedal** choice sic deque sic corr de que corr choice daram

Arquitetura 1: compilação de córpus e criação de glossários



Arquitetura 2: acesso a córpus e criação de verbetes



Conclusões

- O ambiente pode ser facilmente adaptado para uso em projetos semelhantes ao DHPB
- Contribuições:
  - Metodologia para tratamento de córpus históricos
  - Ferramentas (Protew, Protej, Procorph)
  - Recursos (córpus e glossários)

Referências

ATKINS, S.; CLEAR, J.; OSTLER, N. Corpus design criteria. *Journal of Literary and Linguistic Computing*, v. 7, n. 1, 1992.

ARCHER, D., ERNST-GERLACH A., KEMPKEN S., PILZ T., RAYSON P. The identification of spelling variants in English and German historical texts: manual or automatic. In: *Digital Humanities*, 2006, Paris: Sorbonne, 2006. p. 3-5.

FLEXOR, M. H. O. *Abreviaturas*: Manuscritos dos séculos xvi ao xix. 2. ed. [S.l.]: UNESP, 1991. 468 p.

GIOULI, V.; PIPERIDIS, S.. *Corpora and HLT*: Current trends in corpus processing and annotation. Disponível em: <http://www.larflast.bas.bg/balric/eng_files/corpus_deliverable_final.htm>. Acesso em: 25 fev. 2008.

GIUSTI, R.; CANDIDO JR, A.; MUNIZ, M. C. M.; CUCATTO, L. A.; ALUÍSIO, S. M. Automatic detection of spelling variation in historical corpus: An application to build a Brazilian Portuguese spelling variants dictionary. In: *Corpus Linguistics*, 2007, Londres. Corpus Linguistics, 2007.

HIROHASHI, A. S. *Aprendizado de regras de substituição para normatização de textos históricos*. Dissertação (Mestrado) — Instituto de Matemática e Estatística, USP, São Paulo, 2004.

Referências

RAYSON, P., D. ARCHER AND N. SMITH.VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historic corpora, In Proceedings of Corpus Linguistics 2005, vol. 1, no. 1. Birmingham: Birmingham University.

SARDINHA, T. B. Lingüística de Corpus. Barueri, SP: Manole, 2004.

SINCLAIR, J. Preliminary recommendations on Corpus Typology. EAGLES, 1996. Disponível em: <http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/corpustyp.ps.gz>. Acesso em: 16 fev. 2007.

UNIVERSITÉ DE GENÈVE. The ISO 9126 Standard. 2006. Disponível <http://www.issco.unige.ch/ewg95/node1.html>. Acesso em: 14 nov. 2006.

VALE, O. A. ; CANDIDO JUNIOR, A. ; Muniz ; BENGTSON, C. G. ; Cucatto ; ALMEIDA, G. M. B. ; BIDERMAN, M. T. ; Aluísio . Building a large dictionary of abbreviations for named entity recognition in Portuguese historical corpora. In: American Association for Corpus Linguistics, 2008, AACL 2008.

WYNNE, M. (Ed.). Developing Linguistic Corpora: a guide to good practice. Oxford: Oxbow Books, 2005. Disponível em: <http://ahds.ac.uk/linguistic-corpora/>. Acesso em: 23 fev. 2007.