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

Arnaldo Candido Junior

Orientadora:
Sandra Maria Aluísio

Apoio:
Conselho Nacional de Desenvolvimento Científico e Tecnológico

São Carlos - 02 de abril - NILC – ICMC - USP

Roteiro

- Introdução sobre córpus
- O projeto DHPB
- Construção do córpus
- Criação de glossários
- Acesso ao córpus
- Redação de verbetes
- Ambiente para processamento de córpus de Português Histórico

Córpus

"Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, (...) que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise."

(Sardinha, 2002, p. 18)

Córpus (2)

- Diversas tipologias (Atkins et al., 1992; Sinclair, 1996; Giouli e Piperidis., 2002)
- Internacionais:
  - The Bank Of English (524 milhões),
  - FRANTEXT (150 milhões)
- Nacionais:
  - Córpus do NILC (41 milhões)
  - Tycho-Brahe (2 milhões)

Construção de córpus

- Projeto: escolha dos textos do córpus
  - Representatividade, balanceamento, reusabilidade, extensibilidade
- Compilação: coleta dos textos
  - Obtenção de direitos, coleta, limpeza, conversão de formato

Construção de córpus (2)

- Anotação: tratamento de metadados
  - Administrativos, editoriais, analíticos e descritivos (Wynne, 2005)
  - Padrões internacionais (TEI, XCES, ...)
- Uso: diferentes enfoques (língua, conteúdo e mídia)
  - Profissionais: lexicógrafos, lingüistas computacionais, historiadores, desenvolvedores de software

Ferramentas de apoio

- Compilação: reconhecedores óticos
- Anotação
  - Manual: editores XML
  - Automática: etiquetadores morfossintáticos
  - Semi-automática: revisão manual da anotação automática
- Acesso a córpus: concordanceadores
- Extração de conhecimento: sumarizadores

O Projeto DHPB

- Dicionário Histórico do Português do Brasil (Vale et al., 2008)
- Primeiro projeto do gênero
- No século XVI: PB divergia PP (cultura, fauna, flora)
- 1500-1808 (período pré-imprensa)

O Projeto DHPB (2)

- Textos
  - Cartas dos Jesuítas
  - Documentos dos bandeirantes
  - Relatos dos sertanistas
  - Etc.
- Autores: Brasileiros ou Portugueses (que viveram no Brasil por um longo perído)

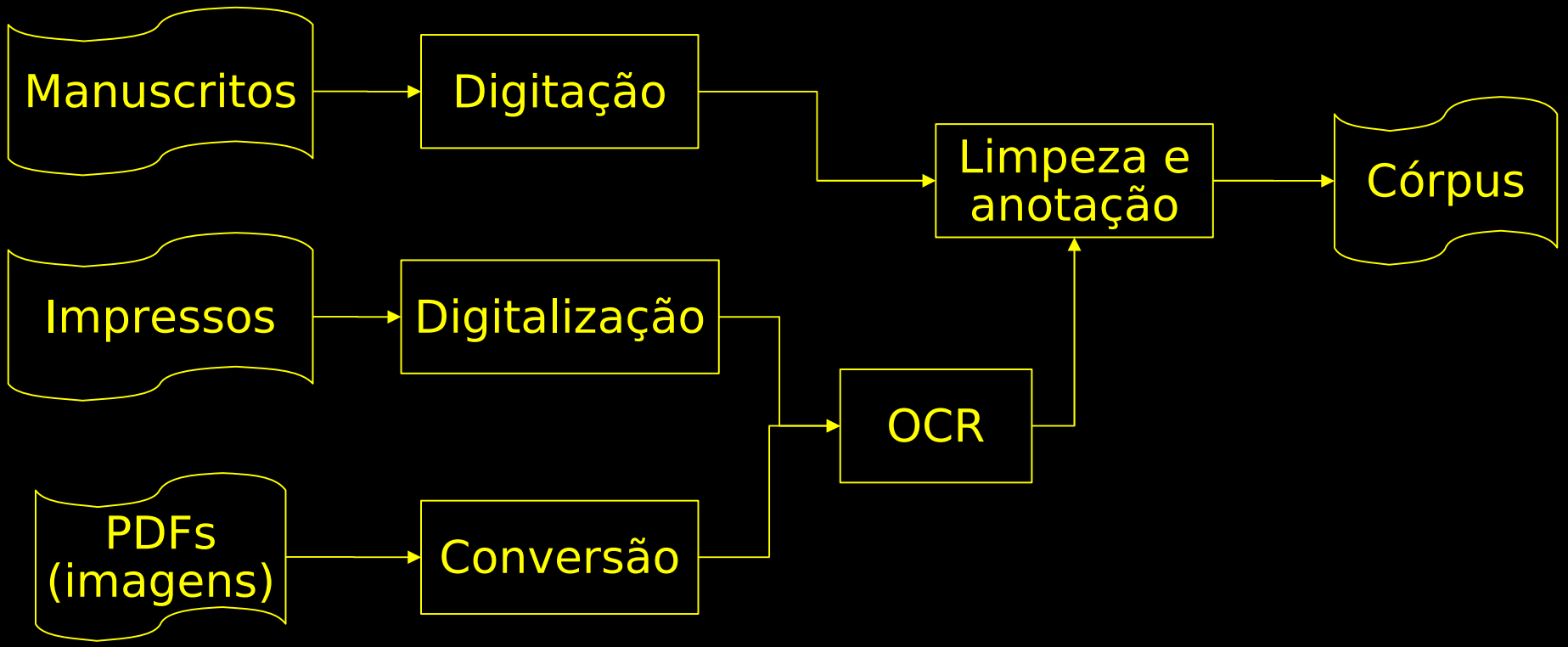
Motivação

- Diversos recursos / metodologias / ferramentas para construção de córpus contemporâneos
  - Poucos atendem às necessidades de córpus históricos
- Diversas ferramentas para dicionários contemporâneos
  - Poucas atendem às necessidades de dicionários históricos

Objetivos

- Levantar as necessidades do projeto DHPB
- Desenvolver metodologias, recursos, e ferramentas para atender estas necessidades
- Criar um ambiente para o processamento de córpus de Português Histórico

Compilação do córpus



Digitalização

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

OCR

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. 205
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 2I0 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. 2I5 Portanto me parece que avião de vir de lá, e, se possivel fosse, com suas molheres e filhos, e alguns que fação taipas e carpinteiros. Cá está hum Mestre para as obras, que hé hum sobrinho[20] de Luis Diaz, mestre das obras d'El-Rey, ho qual veo con trinta mil reis de partido.
(...)

I9 Simão Gonçalves. LEITE I 573.
20 Este «bom oficial», sobrinho de Luís Dias, era Diogo Peres.
LEITE I 22.

Limpeza e anotação

 {7. - BAÍA 9 DE AGOSTO DE 1549 127 - A00_0002.txt,.N}

 convertidos, onde estaremos Vicente Rodriguez e eu, e hum soldado <note place="foot"n="19"> Simão Gonçalves. LEITE I 573. </note> 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 fosse, com suas molheres e filhos, e alguns que fação taipas e carpinteiros. Cá está hum Mestre para as obras, que hé hum sobrinho <note place="foot"n="20"> Este «bom oficial», sobrinho de Luís Dias, era Diogo Peres. LEITE I 22. </note> de Luis Diaz, mestre das obras d'El-Rey, ho qual veo con trinta mil reis de partido. Este nom hé necessario porque abasta ho tio para as obras de S. A.; a este avião de dar o cuidado do nosso collegio; hé bom official. Serão cá muito necessarias pessoas que teção algodão, que há muito, e outros officiaes.
(...)

Ficha catalográfica

Tipologias
1. Tipo da Fonte: EDIÇÃO IMPRESSA
2.1 Domínio Discursivo/Subdomínio Discursivo:
2.2 Gênero/Subgênero:
3a. Tipologia de Assuntos:
3b. Características Sociolingüísticas do Autor:
4. Descrição: CARTAS JESUÍTICAS DISPOSTAS EM ORDEM CRONOLÓGICA, ORGANIZADAS E, QUANDO PRECISO, TRADUZIDAS E ANOTADAS PELO P.e SERAFIM LEITE (1538-1553) - 3 VOLUMES
5: Localização da Obra: UNESP - CAMPUS DE ARARAQUARA
Fonte
6 Nome do Autor do Texto: P. MANUEL DA NÓBREGA
7: Título do Texto: CARTA DO P. MANUEL DA NÓBREGA AO P. SIMÃO RODRIGUES, BAÍA 9 DE AGOSTO 1549
8. Data em que o Texto foi produzido pelo Autor: BAÍA 9 DE AGOSTO 1549
9. Amostra: INTEGRAL
10. Título da Obra: CARTAS DOS PRIMEIROS JESUÍTAS DO BRASIL
11. Editor: SERAFIM LEITE S. J
12. Organizador/Coordenador (coletânea/livro): SERAFIM LEITE S. J
13. Editora: COMISSÃO DO IV CENTENÁRIO DA CIDADE DE SÃO PAULO
14. Local da Edição: SÃO PAULO
(...)

Anotação do cabeçalho

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE TEI.2 SYSTEM "http://docsouth.unc.edu/dtds/teixlite.dtd">
<TEI.2>
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title> CARTA DO P. MANUEL DA NÓBREGA AO P. SIMÃO RODRIGUES, BAÍA 9
DE AGOSTO 1549 </title>
      <author>
        <name> P. MANUEL DA NÓBREGA </name>
        <date> BAÍA 9 DE AGOSTO 1549 </date>
      </author>
(...)
    </titleStmt>
    <sourceDesc>
      <biblStruct>
        <monogr>
(...)
          <imprint>
            <pubDate> 1956</pubDate>
          </imprint>
        </monogr>
      </biblStruct>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

Compilação do córpus

- Fases
  - Codificação do córpus
  - Limpeza
  - Anotação
  - Formatos e versões
- Ferramentas: Protej e Protew
- Estatísticas do córpus compilado

Codificação do córpus

Acentos combinados

Símbolo	Unicode	Freqüência	Exemplo
◌̂	0302	2	quarŷ
◌̃	0303	24.568	comẽrcio
◌̄	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

Limpeza

Formatação, hifenização, numeração de linhas e de parágrafos.

(...) , E mortes que sem se lHe dar Cauza tem **ex= ecutado** no Rio da Madeira o Gentio da Nasçaô **Mu= ra** , impedindo o Comersio dos Moradores naquelle Rio, E pondo em temor, E consternaçaô as Missoenz **esta= blecidas** nelle. Ordeno ao Dout.or Ouvidor geral desta (...)

(...) , E mortes que sem se lHe dar Cauza tem **executado** no Rio da Madeira o Gentio da Nasçaô **Mura**, impedindo o Comersio dos Moradores naquelle Rio,E pondo em temor, E consternaçaô as Missoenz **estabelecidas** nelle. Ordeno ao Dout.or Ouvidor geral desta (...)

Anotação

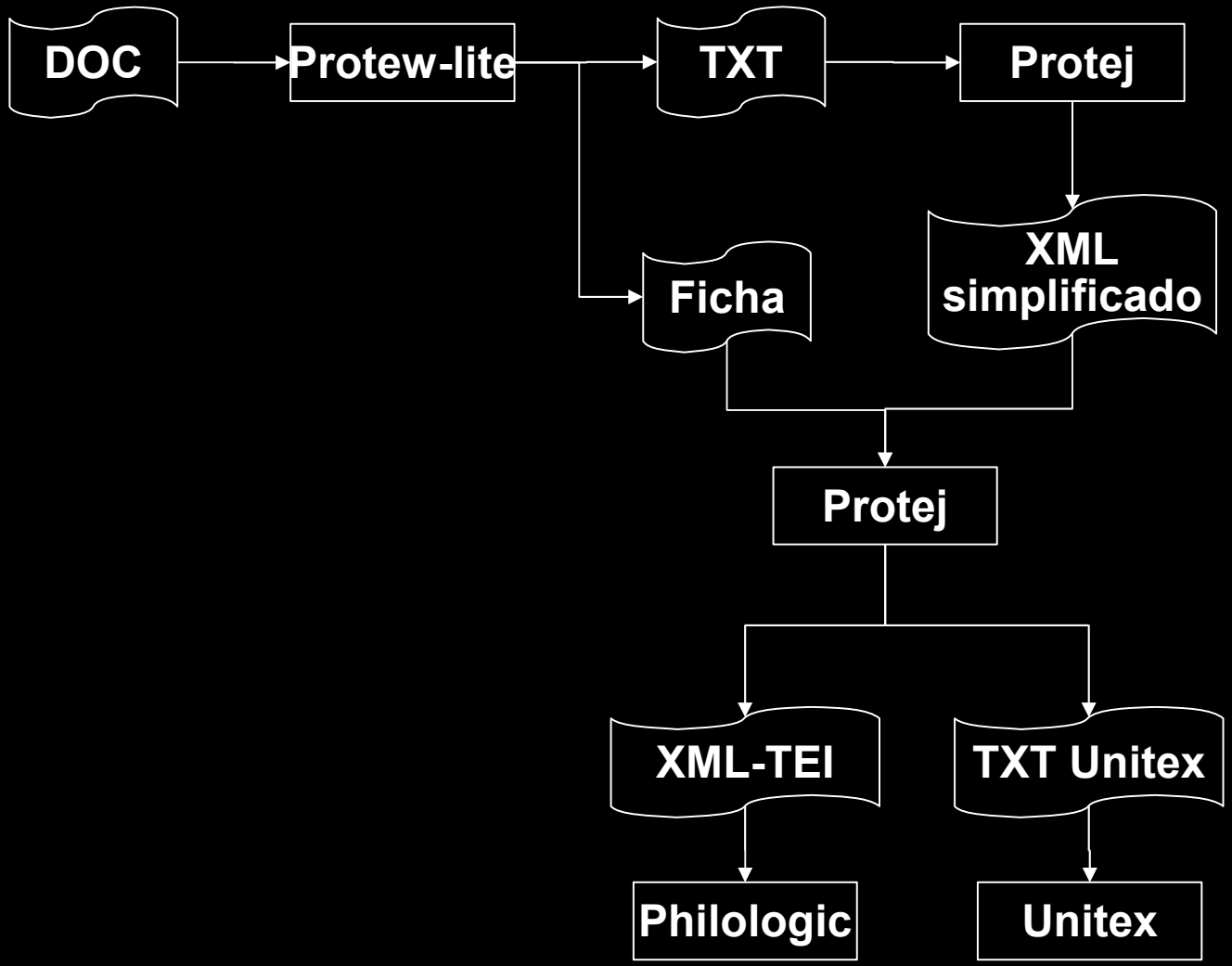
Edição de texto e inserção de etiquetas

(...) da pose vyrem: Como, no anno do nacimento
de Nosso Senhor Jesu Christo de mil e quynhentos
e sessenta 1 anos, haos dose dias do mes d'Aguosto do dyto
(...)
campo e borda do matto, Fernão Jorge juiz hordinario
dyta vila e campo, ante my apareceu ho Irmão (...)

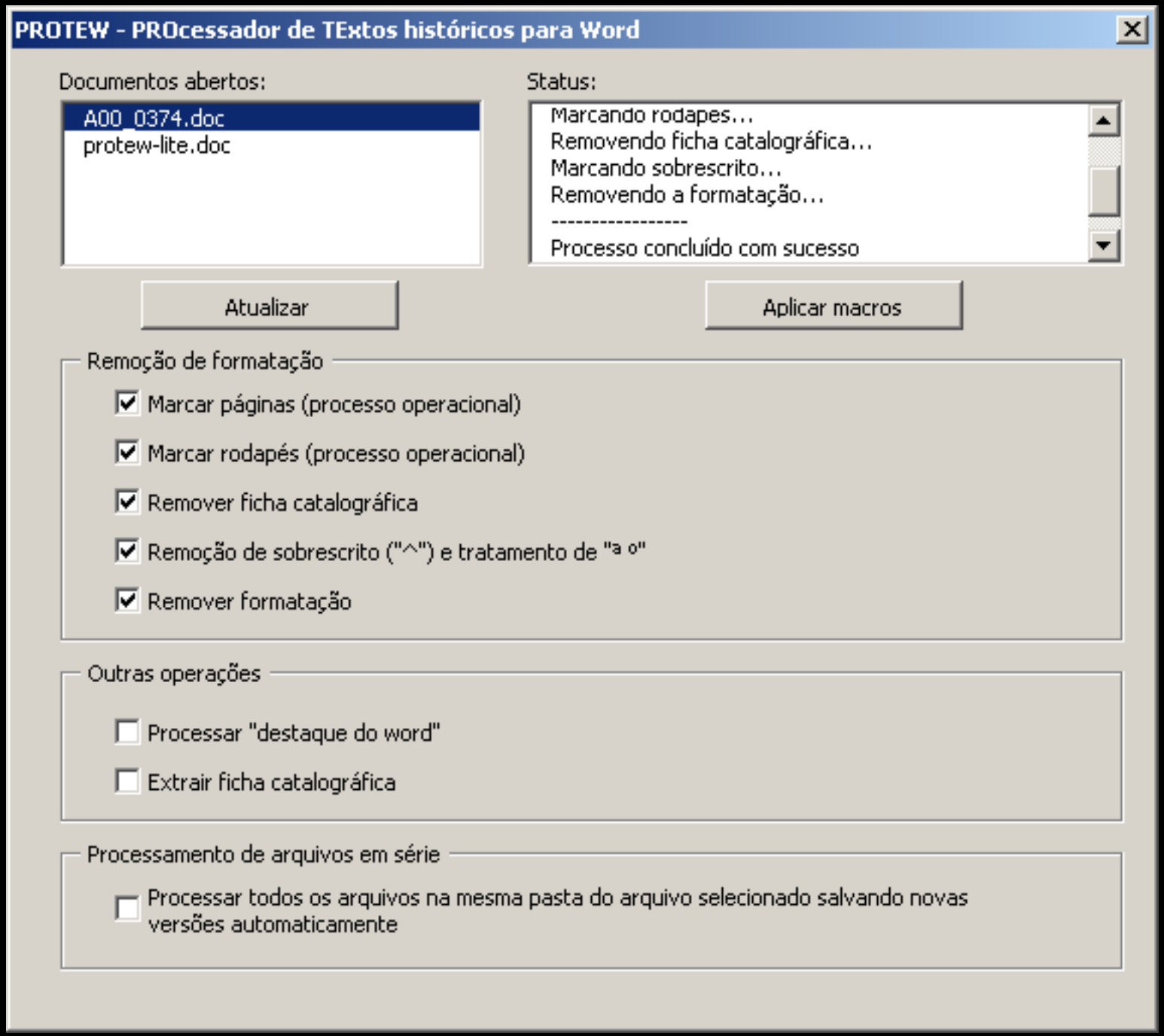
6 **Impresso setenta em vez de sessenta. (...)**

(...) da pose vyrem: Como, no anno do nacimento
de Nosso Senhor Jesu Christo de mil e quynhentos
e sessenta 1 anos, haos dose dias do mes d'Aguosto do dyto
<note place="foot"n="6" type="line">Impresso setenta (...) </note>
(...)
campo e borda do matto, Fernão Jorge juiz hordinario
dyta vila e campo, ante my apareceu ho Irmão (...)

Geração de versões

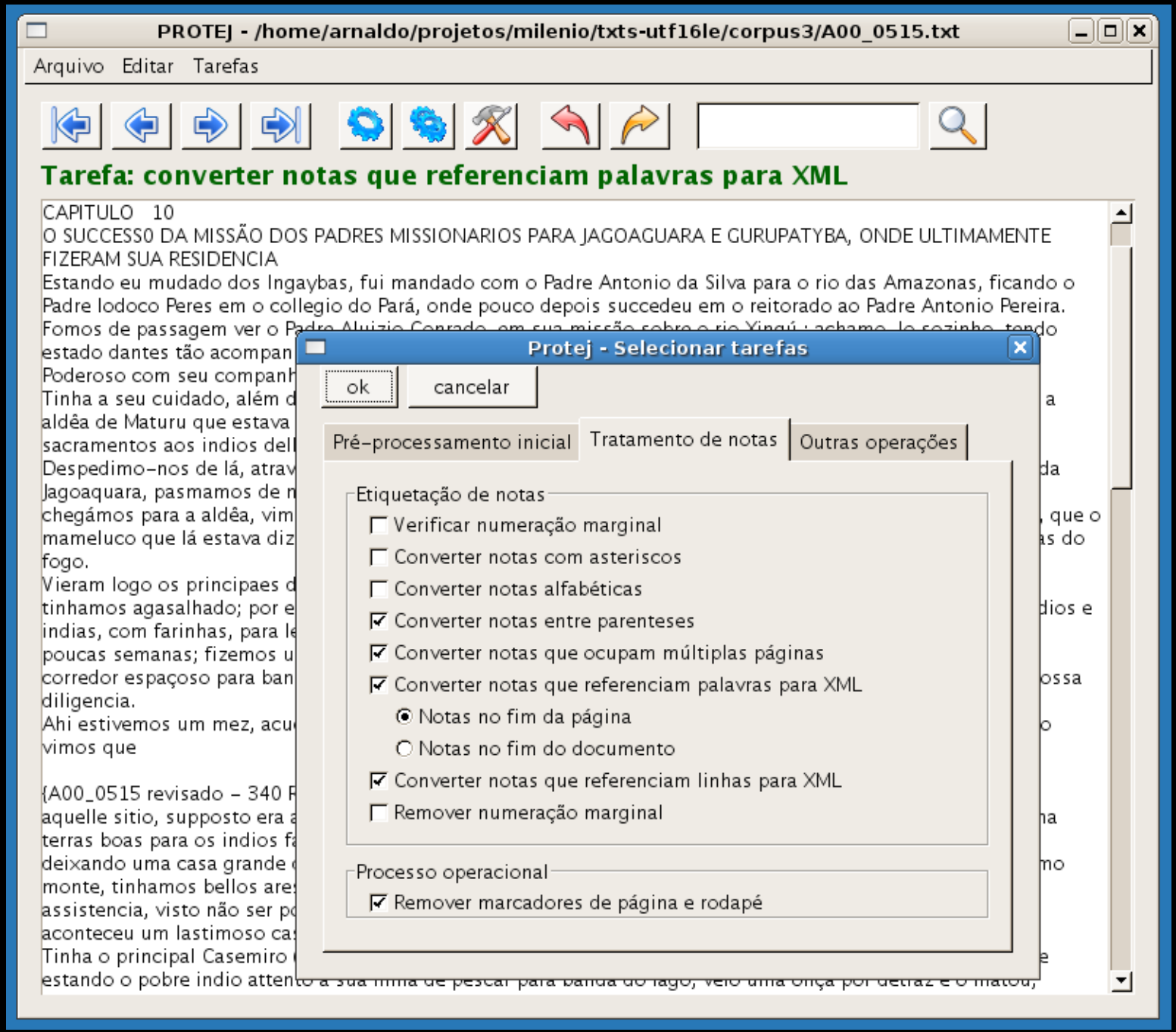


Protew-lite

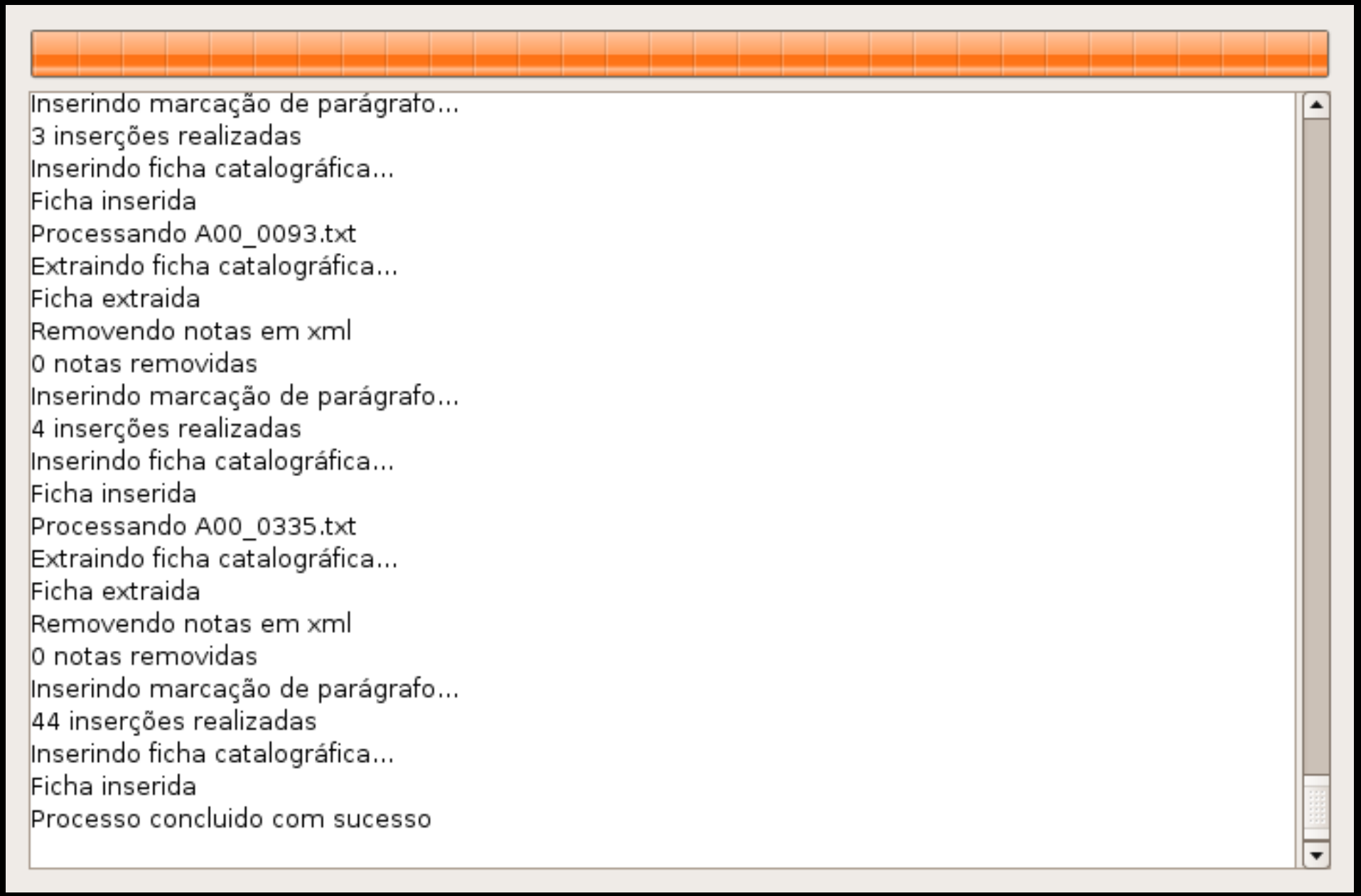

PROTEW - PROcessador de TExtos históricos para Word
Documentos abertos:
A00_0374.doc
protew-lite.doc
Status:
Marcando rodapes...
Removendo ficha catalográfica...
Marcando sobrescrito...
Removendo a formatação...

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

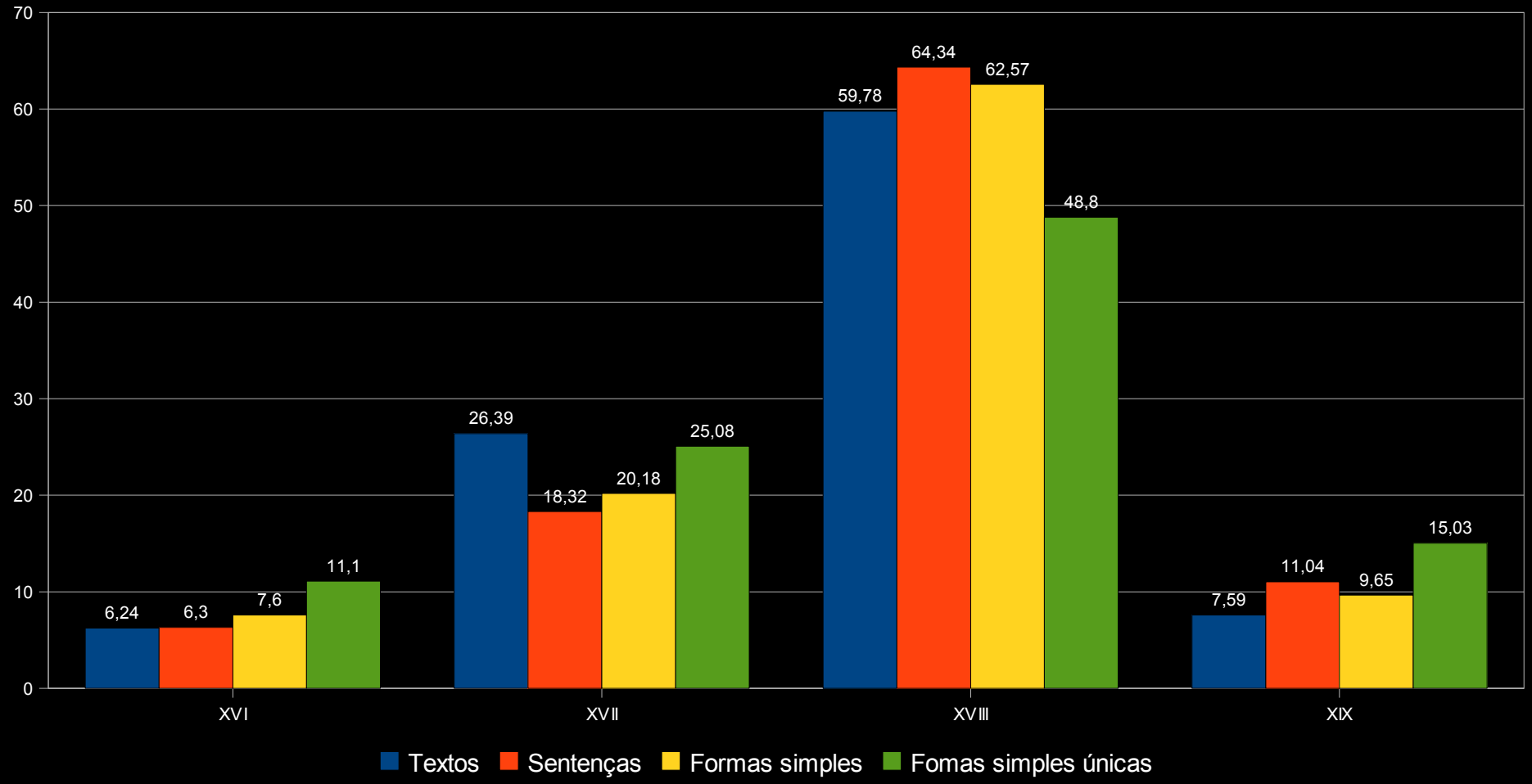
Protej – Córpus Philologic



Dados do córpus

Dados	**Valores**
Tokens	16.505.808
Types	368.850
Formas simples	7.492.473
Formas simples únicas	368.529
Sentenças	287.570
Textos	2.458
Tamanho em MegaBytes (UTF-16)	82

Textos por século (%)



Glossários

- Fenômenos observados: abreviaturas, junções e variantes de grafia
- Dificultam a contagem de freqüências e criação de concordâncias
  - Dificultam a compreensão do texto
- Soluções: substituição, anotação (manual ou automática), criação de glossários

Junções

- Levantadas por uma bolsista do Projeto para separação automática de Junções
- Casos comuns: artigos e preposições
- Exemplos: asmesmas, doestillo, emhumapontadeterra, serraniasque, sobpena, ...
- 10.369 junções

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: extraído de Flexor (1991)
- Glossário C: extraído do córpus
- Protej usado para pré-processamento

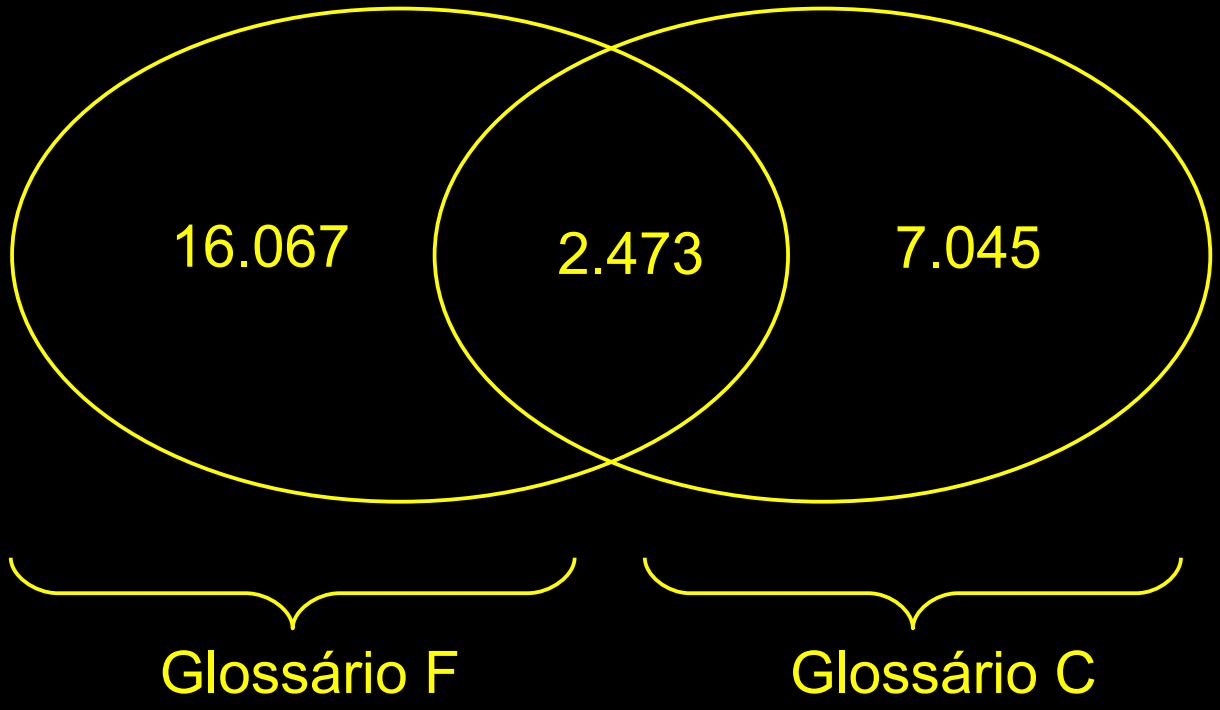
Glossário F

- Convertido para formato digital por uma bolsista do projeto
- Anotação morfossintática (A, B e C)
- Anotação semântica para reconhecimento de entidades nomeadas (A, B, e C)
- 21.869 abreviaturas com 8.721 expansões
- 16.067 abreviaturas simples e 5.635 expansões
- Abreviaturas simples no córpus: 18,92%

Glossário C

- Apenas abreviaturas (sem expansões)
- 7.045 abreviaturas
  - Presença de sobrescrito: ant.^o, cid.^e, 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%)

Glossário C versus glossário F



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

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

Variantes de grafia

- Inexistência de um sistema ortográfico em textos históricos
- Exemplo: villa, vyla, vjlla, vylla, vjla
- Diversas abordagens: Hirohashi (2004), Rayson et al. (2005), Archer et al. (2005)
- Giusti et al. (2007): regras de transformação (Siaconf) e distância de edição (Philologic + Agrep)

Regras de transformação

- Triplas (E_1 E_2 S), 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

Philologic x Siaconf

Técnica	Verdadeiros Positivos	Falsos Positivos	Precisão	Cobertura Comparativa
Regras de transformação (siaconf)	36	0	100%	72%
Distância de edição (Philologic)	41	196	20.92%	84%

Fenômenos combinados

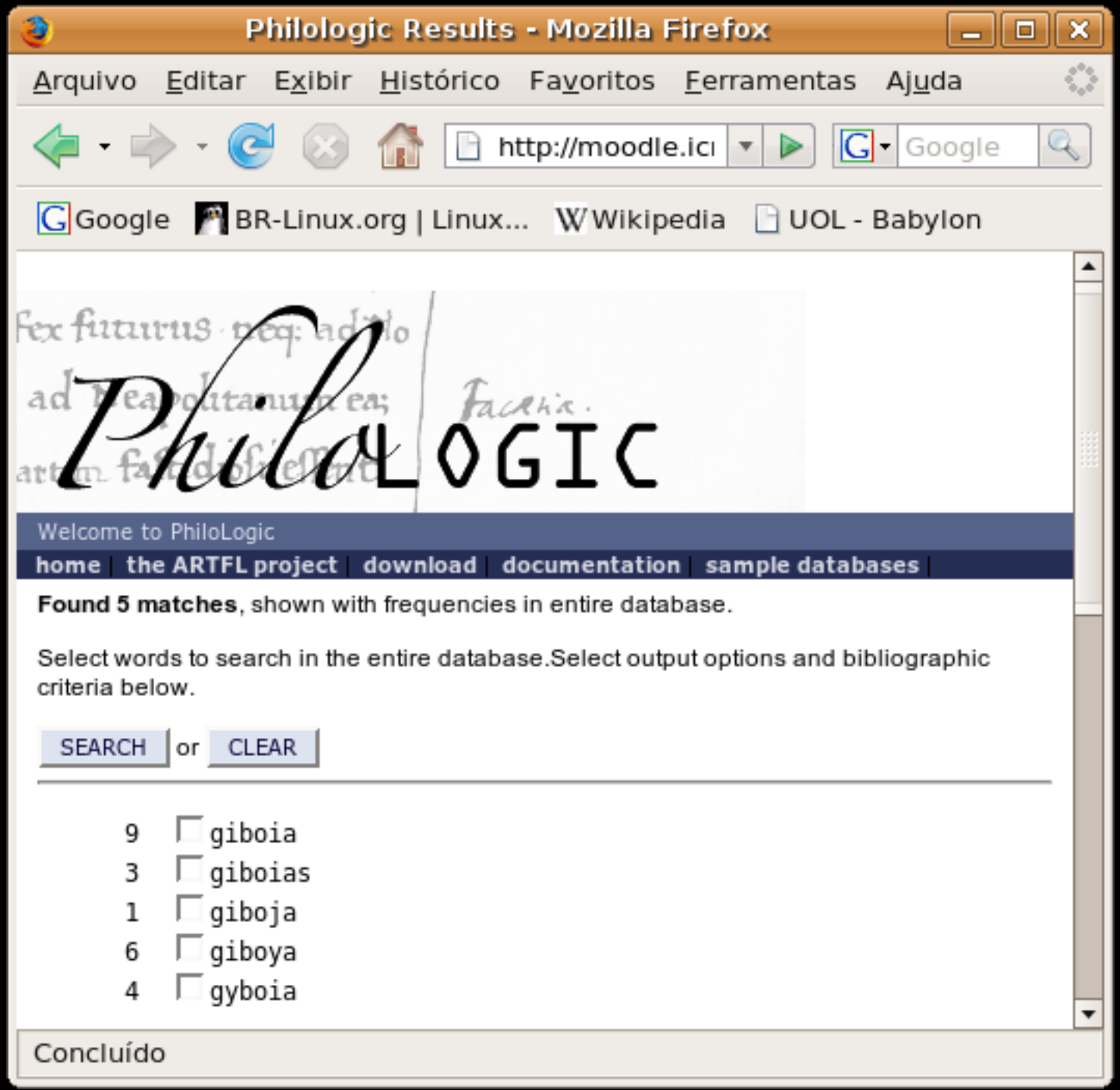
- Não foram considerados

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

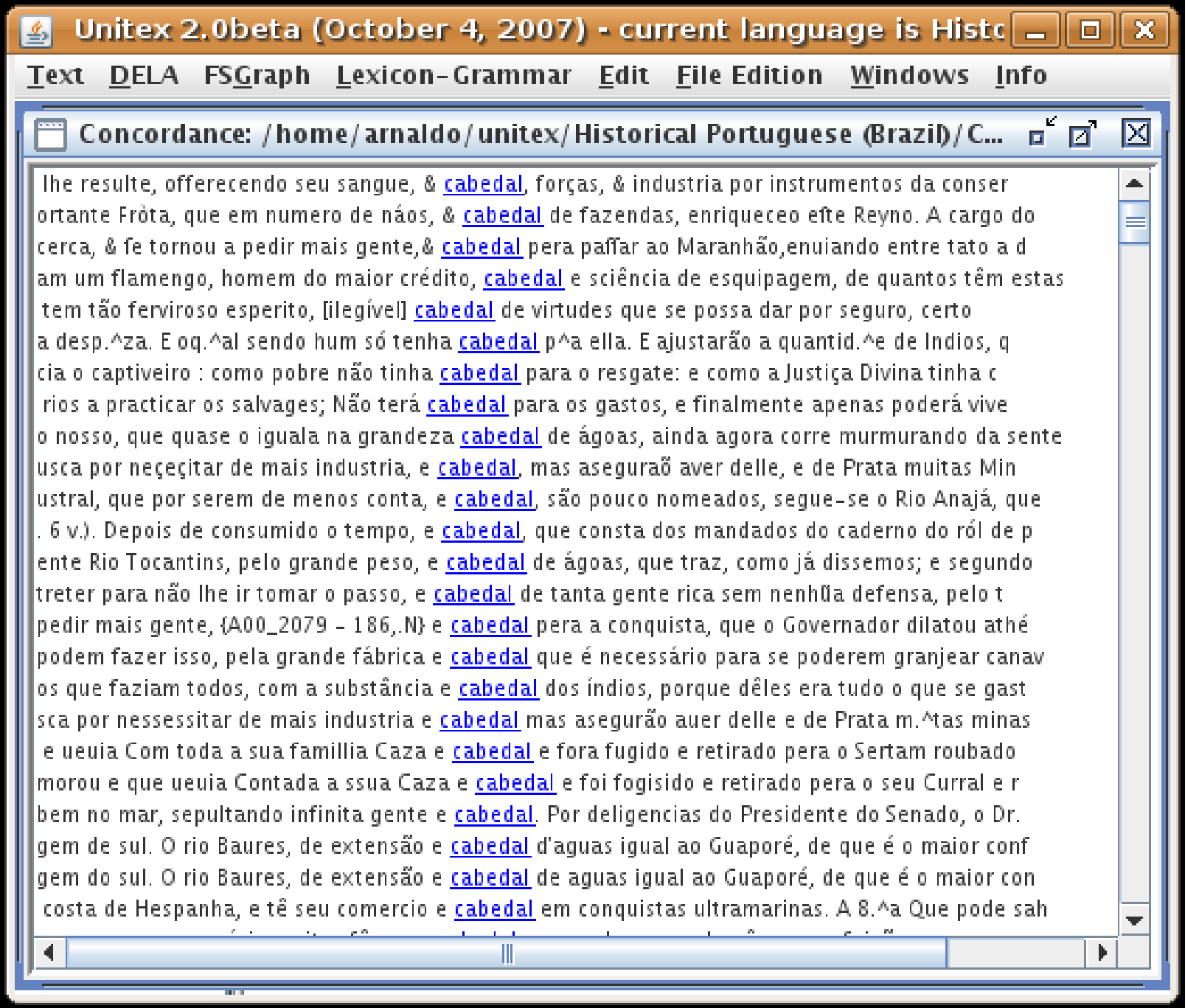
Acesso a córpus

- Processadores analisados
  - Philologic: TEI + Interface Web + variantes
  - Unitex: glossários DELA
  - GATE
  - Tenka Text
  - Xaira
- Critério: ISO 9126 (UNIVERSITÉ DE GENÈVE, 2006)

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

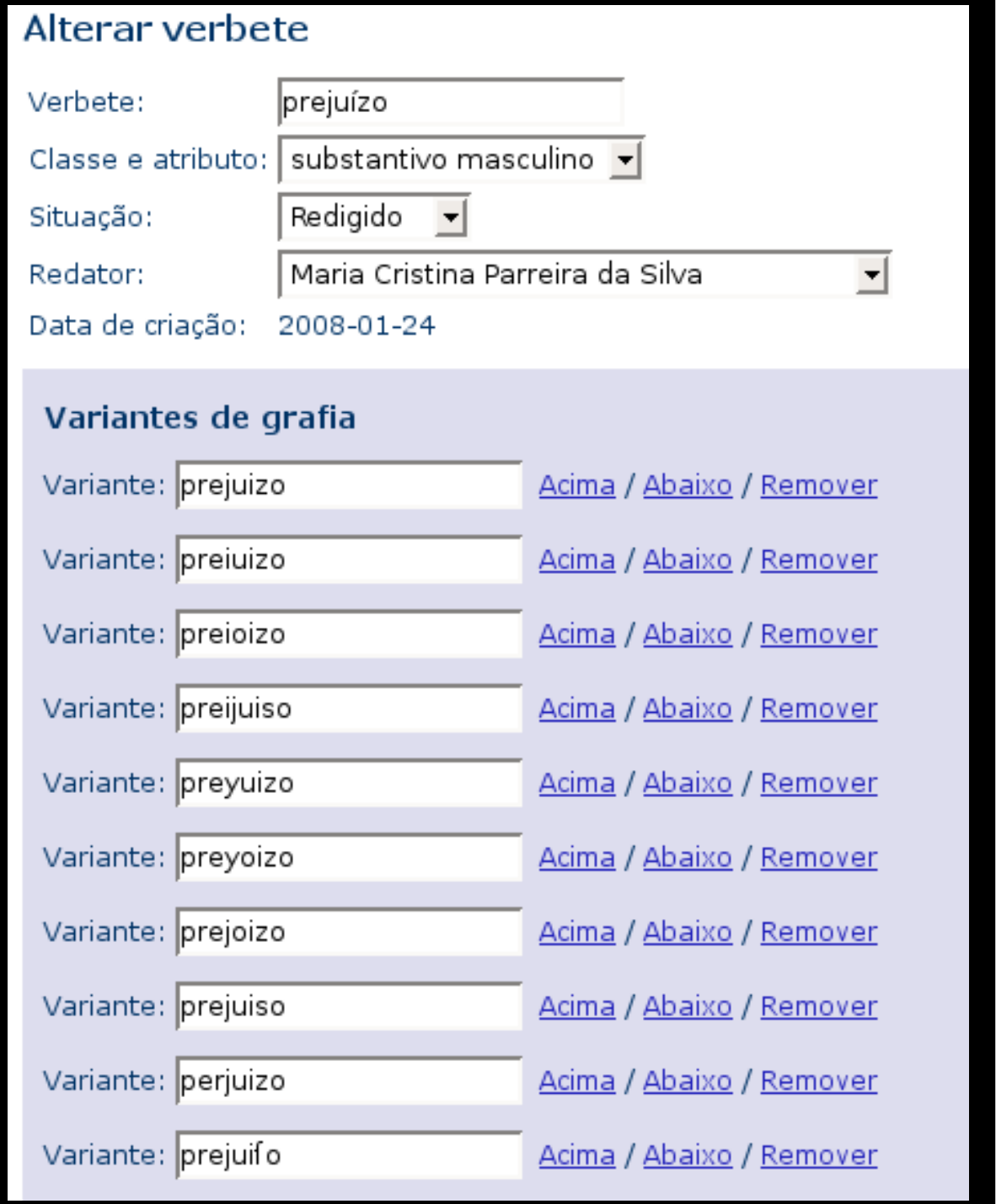

Unitex 2.0beta (October 4, 2007) - current language is Histc
Text DELA FSGraph Lexicon-Grammar Edit File Edition Windows Info
Concordance: /home/arnaldo/unitex/Historical Portuguese (Brazil)/C...
lhe resulte, offerecendo seu sangue, & cabedal, forças, & industria por instrumentos da conser
ortante Frôta, que em numero de náos, & cabedal de fazendas, enriqueceo eſte Reyno. A cargo do
cerca, & ſe tornou a pedir mais gente,& cabedal pera paſſar ao Maranhão,enuiando entre tato a d
am um flamengo, homem do maior crédito, cabedal e sciência de esquipagem, de quantos têm estas
tem tão ferviroso esperito, [ilegível] cabedal de virtudes que se possa dar por seguro, certo
a desp.^za. E oq.^al sendo hum só tenha cabedal p^a ella. E ajustarão a quantid.^e de Indios, q
cia o captiveiro : como pobre não tinha cabedal para o resgate: e como a Justiça Divina tinha c
rios a practicar os salvages; Não terá cabedal para os gastos, e finalmente apenas poderá vive
o nosso, que quase o iguala na grandeza cabedal de ágoas, ainda agora corre murmurando da sente
usca por neçeçitar de mais industria, e cabedal, mas aseguraõ aver delle, e de Prata muitas Min
ustral, que por serem de menos conta, e cabedal, são pouco nomeados, segue-se o Rio Anajá, que
. 6 v.). Depois de consumido o tempo, e cabedal, que consta dos mandados do caderno do ról de p
ente Rio Tocantins, pelo grande peso, e cabedal de ágoas, que traz, como já dissemos; e segundo
treter para não lhe ir tomar o passo, e cabedal de tanta gente rica sem nenhũa defensa, pelo t
pedir mais gente, {A00_2079 - 186,.N} e cabedal pera a conquista, que o Governador dilatou athé
podem fazer isso, pela grande fábrica e cabedal que é necessário para se poderem granjear canav
os que faziam todos, com a substância e cabedal dos índios, porque dêles era tudo o que se gast
sca por neſſeſſitar de mais industria e cabedal mas asegurão auer delle e de Prata m.^tas minas
e ueuia Com toda a sua famillia Caza e cabedal e fora fugido e retirado pera o Sertam roubado
morou e que ueuia Contada a ssua Caza e cabedal e foi fogisido e retirado pera o seu Curral e r
bem no mar, sepultando infinita gente e cabedal. Por deligencias do Presidente do Senado, o Dr.
gem de sul. O rio Baures, de extensão e cabedal d'aguas igual ao Guaporé, de que é o maior conf
gem do sul. O rio Baures, de extensão e cabedal de aguas igual ao Guaporé, de que é o maior con
costa de Hespanha, e tê seu comercio e cabedal em conquistas ultramarinas. A 8.^a Que pode sah

Suporte a edição de verbetes

- Procorph: edição e formatação de verbetes
  - Variantes de grafia
  - Datação
  - Referências ao córpus
  - Acepções (definições)
  - Abonações (excertos do córpus)

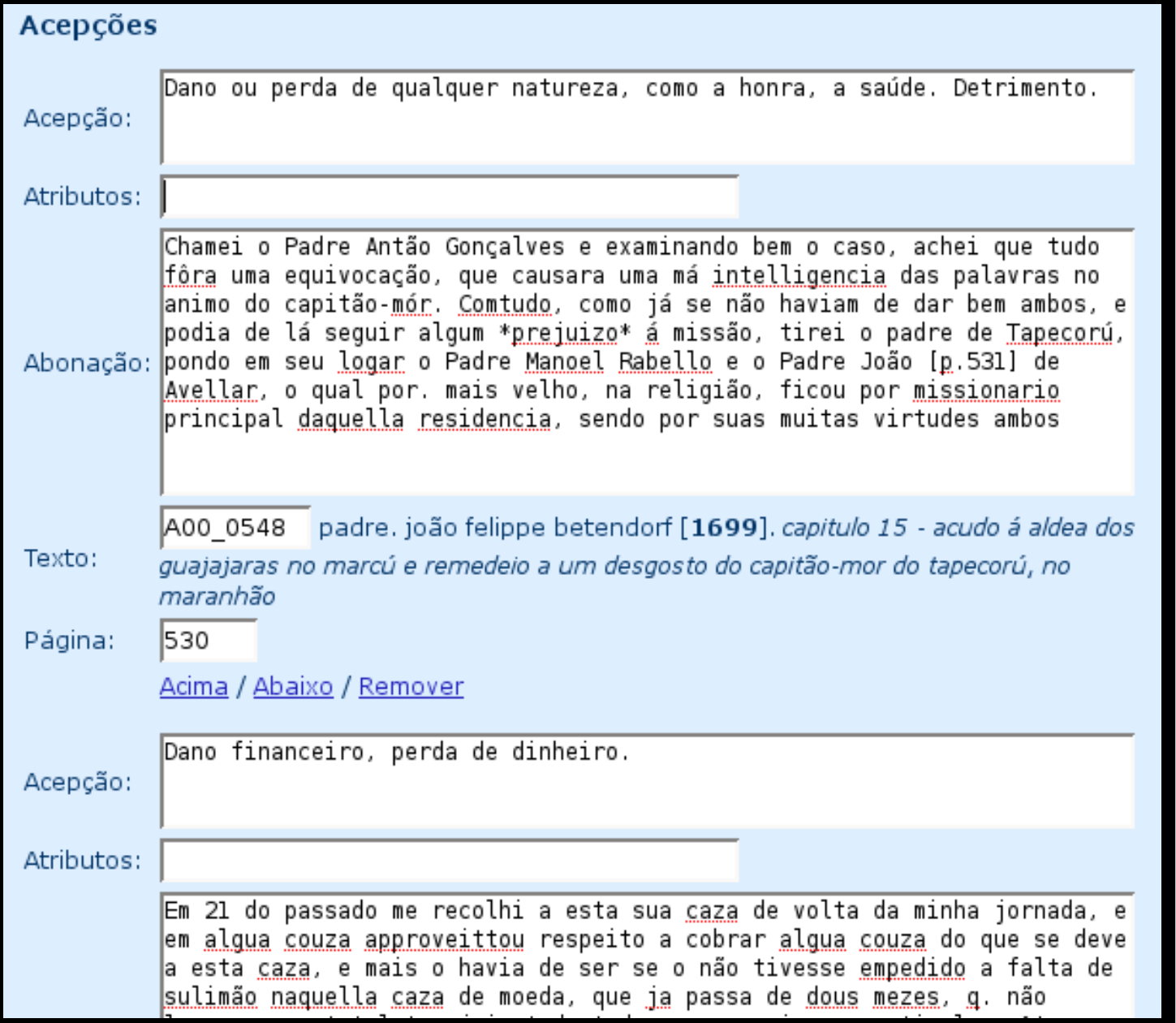
Procorph

- Dados gerais
- Variantes



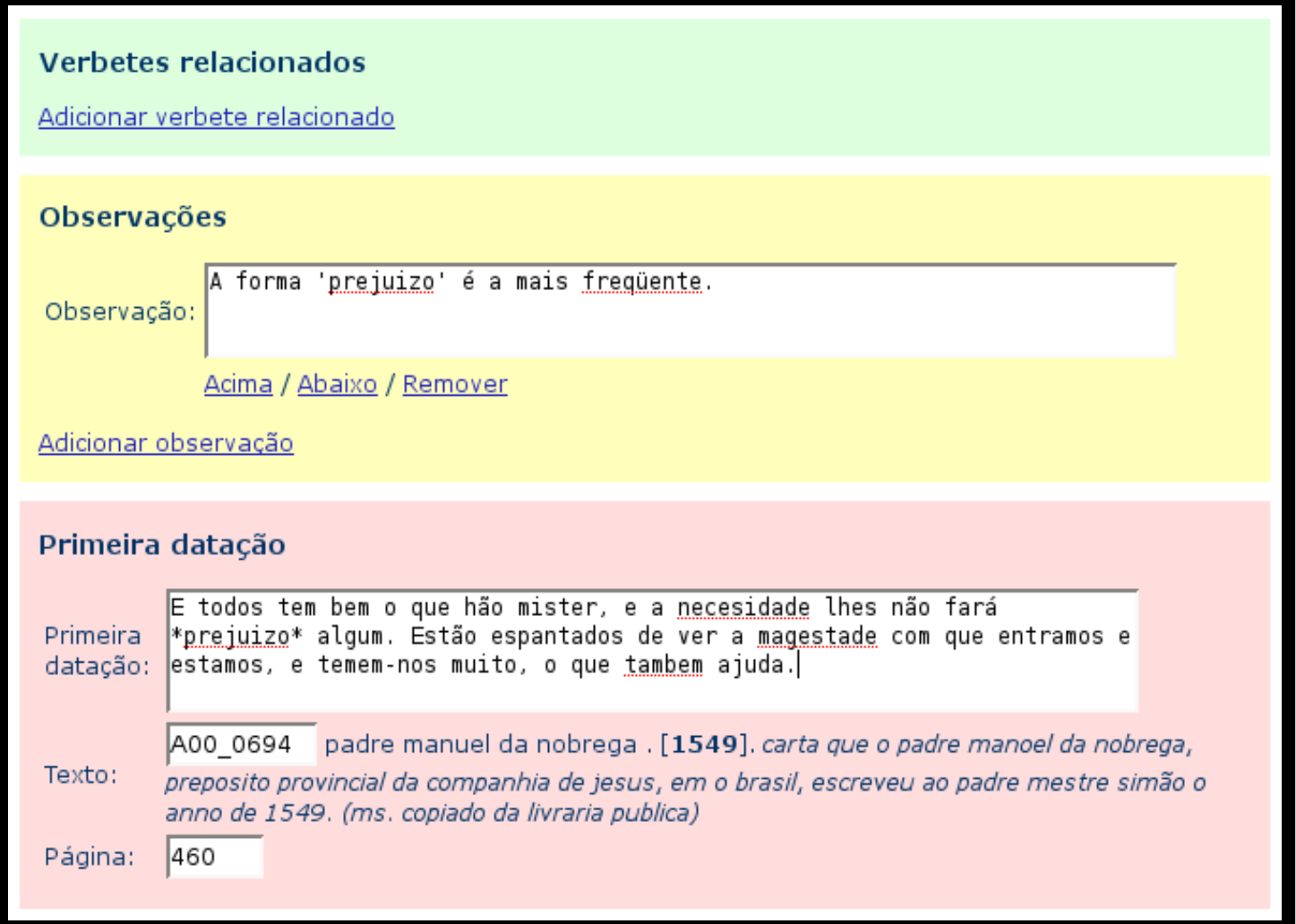
Procorph

- 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.

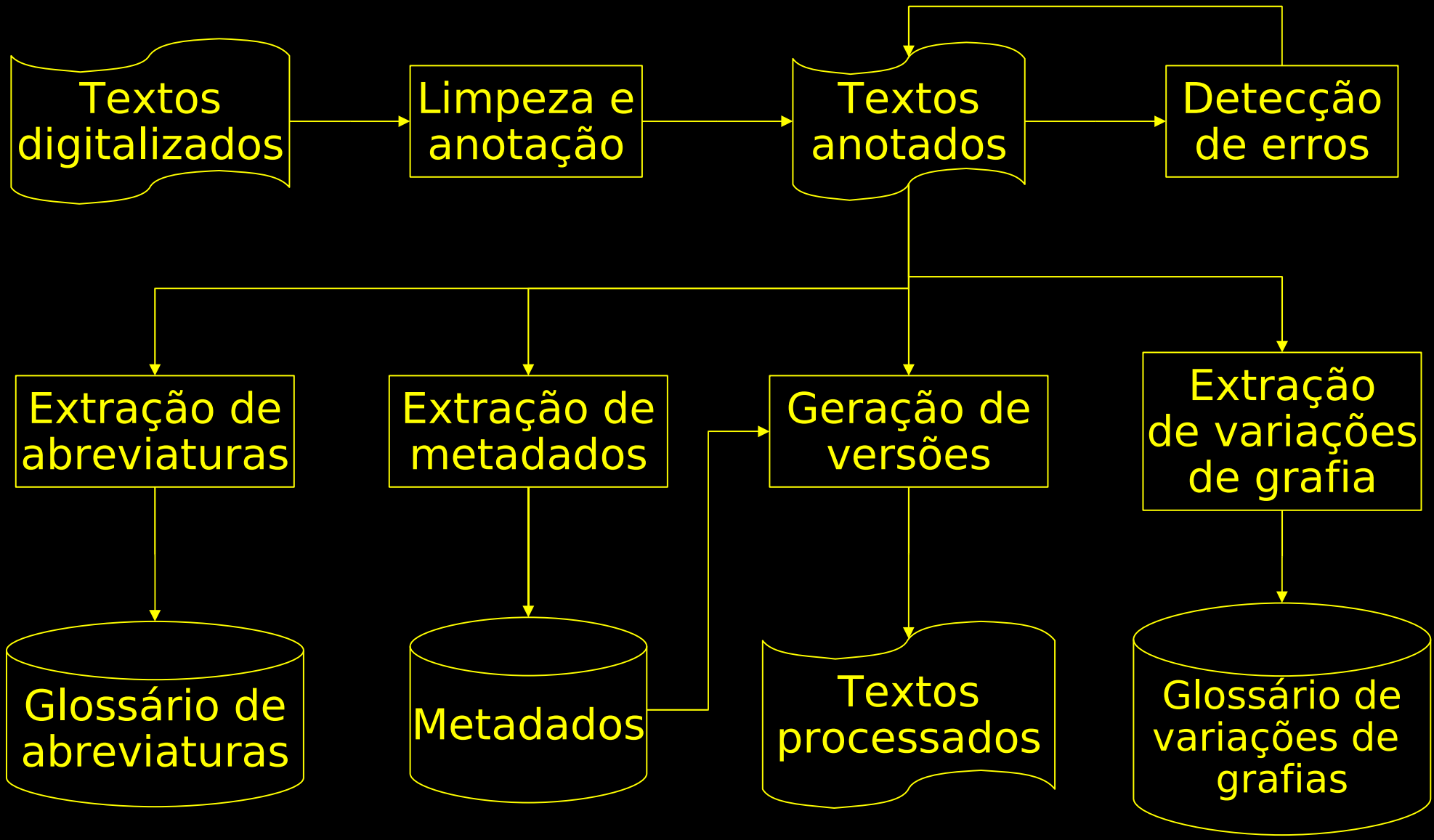
> Porque quem vos mostrara, há hoje trezentos anos, uma cana de que se faz o **açúcar** e vos dissera que daquela cana se havia de formar com a indústria humana um pão de **açúcar** tão formoso como hoje o vemos, tê-lo-ieis por cousa ridiculosa; e, pelo conseguinte, se vos fôsse mostrado um pedaço de 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**.

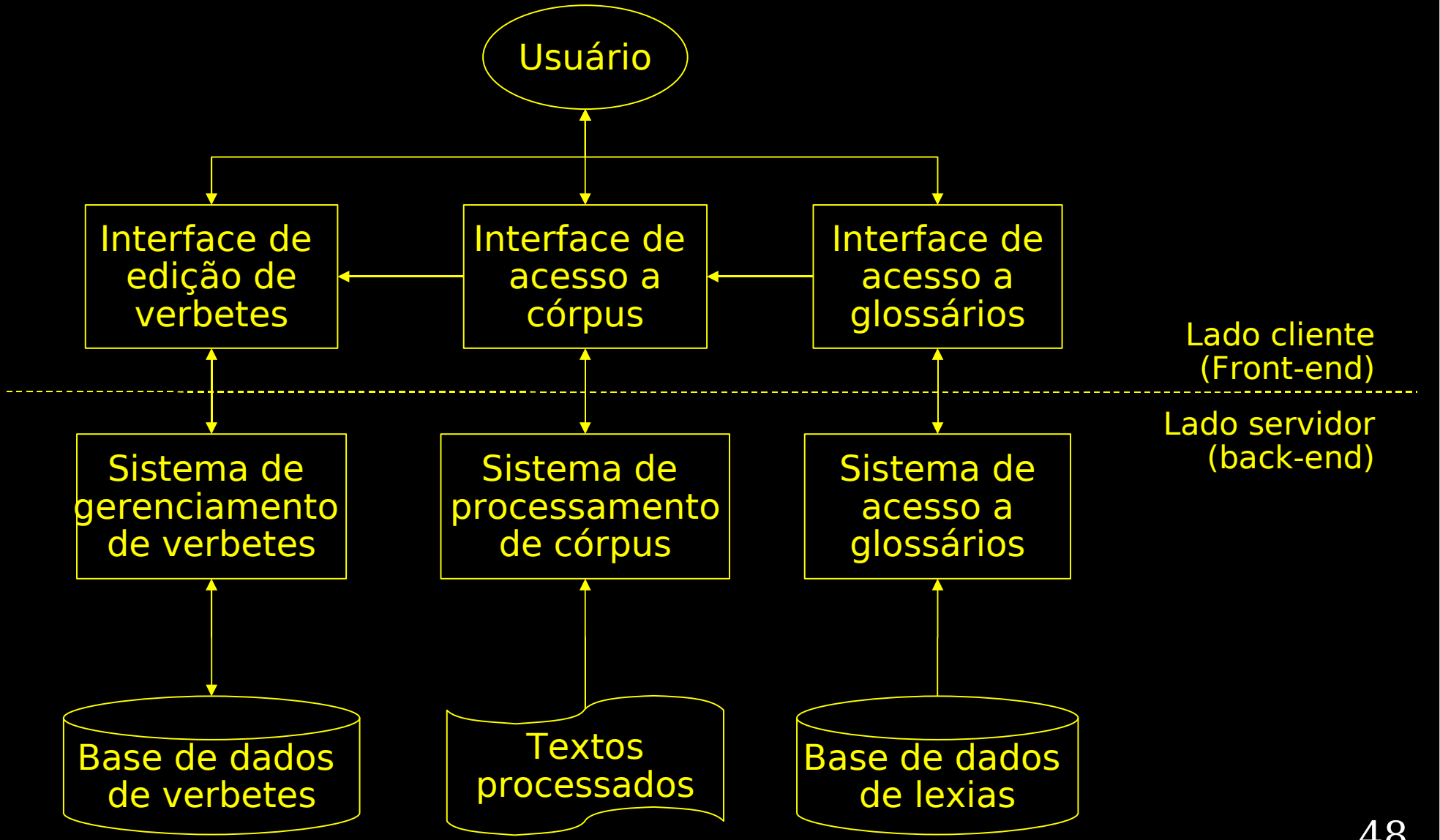
Ambiente para processamento de córpus de Português Histórico

- Arquitetura 1:
  - Compilação do córpus
  - Construção de glossários
  - Geração de versões
- Arquitetura 2:
  - Acesso a córpus
  - Acesso a glossários
  - Edição de verbetes

Arquitetura 1



Arquitetura 2



Trabalhos futuros

- Separação automática de junções
- Extração dos metadados domínio e gênero
- Detecção de erros de OCR
- Novas etiquetas TEI
- Novas regras de transformação
- Acesso a glossários no Procorph
- Acesso a córpus no Procorph

Conclusões

- Este trabalho faz parte do projeto DHPB
- Contribuições:
  - Metodologia
  - Ferramentas (Protew, Protej, Procorph)
  - Recursos (córpus e glossários)
- As contribuições podem ser usadas em outros projetos

Referências citadas na apresentação

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 citadas na apresentação

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.