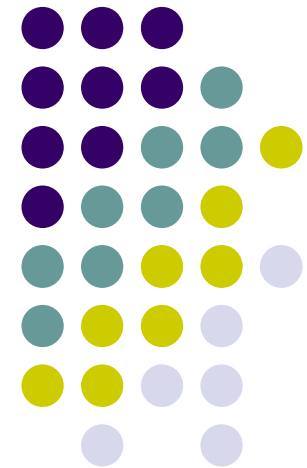


*Córpus Históricos,
Recursos Léxicos e Ferramentas
para a tarefa de criação de dicionários*

*Sandra Maria Aluísio
NILC-ICMC-USP*



I Escola Brasileira de Linguística Computacional
3 a 5 de setembro de 2007



Agenda



- 1) O projeto Dicionário Histórico do Português do Brasil (DHPB)
- 2) Desafios na construção de corpúscos históricos
 - tipologia de textos
 - anotação dos metadados e dos textos
 - codificação de caracteres que caíram em desuso
 - abreviaturas
 - variação de grafia
 - junção das palavras
- 3) Processadores de corpúscos mais adaptados corpúscos históricos
 - O Philologic e o Unitex - uma análise comparativa.
- 4) Tarefas no contexto de criação de verbetes para um dicionário histórico
 - uso do Unitex e do Philologic;
 - uso dos dicionários de variação de grafia e do PB contemporâneo



Projeto DHPB

- Projeto do programa Institutos do Milênio (CNPq)
- Duração de 3 anos (2006-2008)
- Instituição-sede:
 - FCL da UNESP, Araraquara (coordenadora Profa. M. T. Biderman)
- Instituições parceiras:
 - Universidade de Évora,
 - Universidade de São Paulo, Campus de São Paulo e Campus de São Carlos,
 - Universidade Federal de São Carlos,
 - Universidade Federal de Mato Grosso do Sul,
 - Universidade Federal do Rio Grande do Sul,
 - Faculdade de São Bento (Mosteiro de São Bento, Bahia),
 - Universidade Federal de Minas Gerais,
 - Universidade Federal de Uberlândia,
 - Universidade Federal da Bahia
 - Universidade Estadual de Londrina,
 - Escolas Integradas Nossa Sra. da Ressurreição (Catanduva/SP),
 - UNESP, S. J. Rio Preto
- 21 pesquisadores seniores, alunos de graduação e mestrado

Objetivo do Projeto DHPB



- Preenchimento de uma lacuna na cultura brasileira:

“O projeto pretende dotar os brasileiros com um **dicionário** que analisará e descreverá o vocabulário do Português Brasileiro em seu período de formação, ou seja, nos séculos **XVI, XVII e XVIII**, quando a língua do Brasil ainda era caudatária do Português Europeu, porém, já ia armazenando um vocabulário forjado em nossas plagas.”

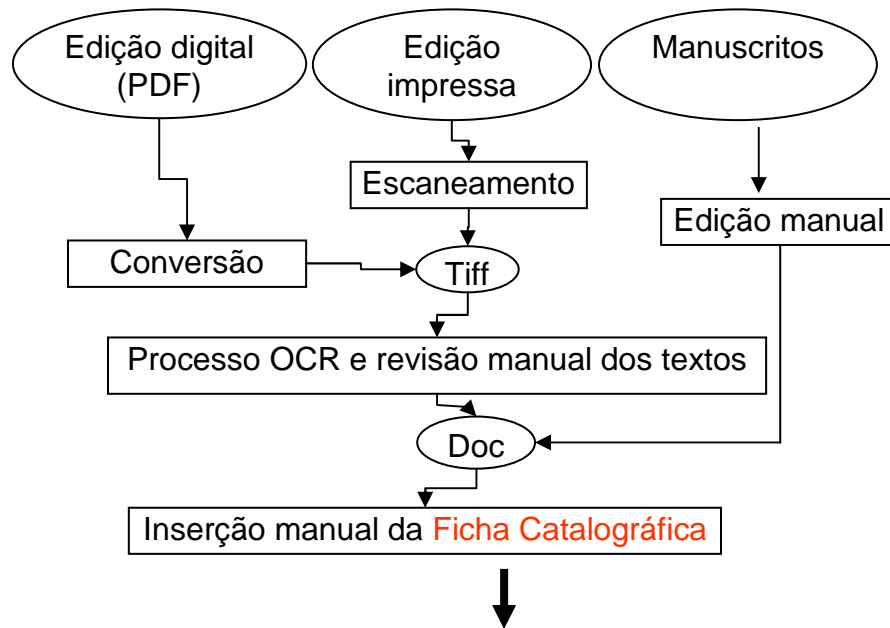
(Biderman, projeto)

Cópus do DHPB



- Textos de 1500-1808 (vinda da família real portuguesa; período pré-imprensa)
 - Textos sobre o Brasil e produzidos por brasileiros, ou portugueses radicados definitivamente no país
 - para permitir a recuperação do repertório vocabular usado nos séculos XVI, XVII e XVIII.
 - Tipos de Texto
 - Cartas dos Jesuítas
 - Documentos dos bandeirantes
 - Relatos dos sertanistas, etc.
 - Tipos de Fontes
 - Manuscritos: editados manualmente
 - Documentos impressos publicados: escaneamento, OCR, revisão manual dos textos
 - Documentos digitais (PDF): conversão de formato, OCR, revisão manual dos textos

Pré-processamento



- Padrão Dublin Core
- Metadados de outros projetos de Córpus Históricos
- Experiência com metadados de Córpus Contemporâneas
- Necessidades do projeto DHPB

Tipologias	
1. Tipo da Fonte:	EDIÇÃO IMPRESSA
2.1 Domínio Discursivo/Subdomínio Discursivo:	CIENTÍFICO
2.2 Gênero/Subgênero:	
3a. Tipologia de Assuntos:	
3b. Características Sociolingüísticas do Autor:	FREIMANDEL DA CONCEIÇÃO VELLOSO ERA FRANCISCANO, NASCIDO EM 1742 NA VILA DE SÃO JOSÉ, COMARCA DO RIO DAS MORTES, EM MNAS GERAIS. ESTUDOU FILOSOFIA E TEOLOGIA, E DEDICOU-SE AO ESTUDO DA BOTÂNICA. MORREU EM 1811 NO CONVENTO DE SANTO ANTÔNIO, NO RIO DE JANEIRO.
4. Descrição:	O FAZENDEIRO DO BRASIL CULTIVADOR, EM SEU TOMO III, PARTE III, TRAZ INFORMAÇÕES SOBRE AS BEBIDAS ALIMENTOSAS E EM ESPECIAL SOBRE O CACAU. A OBRA ESTÁ DIVIDIDA EM 3 PARTES E EM CAPÍTULOS; AO FINAL, HÁ A MEMÓRIA SOBRE O CACAU E CHOCOLATE.
5: Localização da Obra:	INSTITUTO DE ESTUDOS BRASILEIROS
Fonte	
6 Nome do Autor do Texto:	FREIMARIANO DA CONCEIÇÃO VELLOSO
7: Título do Texto:	O FAZENDEIRO DO BRAZIL, CULTIVADOR. BEBIDAS ALIMENTOSAS, CACAO.
8. Data em que o Texto foi produzido pelo Autor:	ENTRE 1798 E 1806
9. Amostra:	INTEGRAL
10. Título da Obra:	O FAZENDEIRO DO BRAZIL, CULTIVADOR.
11. Editor:	-
12. Organizador/Coordenador (coletânea/livro):	
13. Editora:	IMPRESA RÉGIA
14. Local da Edição:	LISBOA
15: Data da Edição:	1805
16: Número da Edição:	-
17: Volume:	TOMO III –PARTE III
18: Tipografia:	48%
19: Número de Páginas da Obra:	364
20: Número de Páginas Escaneadas da Obra:	170
21: Número de Páginas do Texto:	11 P., SEM NUMERAÇÃO
22: Identificador (ISBN/ISSN/DOI):	
Revisão	
23: Revisor(a) (responsável pela revisão da digitalização):	ROSANE MALUSÁ GONÇALVES PERUCHI
Formato do Arquivo Final (txt)	
24: Codificação:	
25: Data da Integração do Arquivo de Texto ao Córpus:	
26: Tamanho do Texto:	calculado automaticamente

Cópus do DHPB



- Faz uso de textos publicados, com intervenções de editores
 - completaram palavras com rasuras, inseriram notas explicativas
- Intervenções do projeto
 - juntar palavras hifenizadas e
 - separar a junção de palavras
 - por exemplo damesma, agrande, comqualquer ficam da mesma, a grande, com qualquer

“ o nosso foco como lexicógrafos não é o do foneticista/fonólogo nem mesmo o do sintaticista, para os quais a versão *ipsis litteris*, especialmente para o primeiro, é de crucial importância. De fato, o nosso foco principal será a semântica das palavras e do texto.”

(Biderman, relatos de reuniões de projeto)

Córpus do DHPB



- Estimativa do tamanho final do córpus (fim de setembro)
 - Por volta de 2.500 textos e 7 milhões de palavras
- Córpus de trabalho já processado para trabalhar com Unitex e o Philologic
 - 1.733 textos, 4.9 milhão de palavras



Córpus do DHPB já processado

Dados	Séculos			
	XVI	XVII	XVIII	XIX
% Textos	7,22%	23,28%	60,36%	9,13%
% Sentenças (aproximado)	9,07%	23,74%	52,96%	14,23%
% Palavras	9,83%	24,38%	52,97%	12,81%

Distribuição dos Textos por Séculos

Visão do papel do *Córpus* mudou durante o Projeto



- **Função do *córpus* é identificar o texto de onde se extrairá a abonação para o significado/abonação do vocábulo cujo valor semântico/uso contextual será registrado**

Para podermos ter uma base textual informatizada de dimensões relativamente grande é preciso planejar a informatização para o período de **um ano**.

(Biderman, projeto)

“Por outro lado, **concluimos também que a criação do *corpus* informatizado** que estamos gerando e construindo tem uma importância vital para as pesquisas sobre o Português do Brasil e para a história da nossa cultura e da nossa sociedade, **valor esse quase tão grande quanto o próprio dicionário que vamos produzir.**”

(Biderman, relatório parcial do projeto, após 1 ano)



Compilação e processamento do *córpus*: 1 ano e 9 meses



Estágios da compilação de um corpus

- Projeto do corpus, que inclui a seleção dos textos e os cuidados com os requisitos como
 - autenticidade, representatividade, balanceamento, amostragem, diversidade, tamanho e reusabilidade
- Compilação (ou coleta) e conversão de formato
 - Obtenção de direitos de uso
 - Coleta de textos: digitalização, digitação e transcrição
 - Nomeação dos arquivos de textos
 - Limpeza: remoção de dados pessoais e de metadados indesejados
- Anotação **estrutural** (marcação de dados externos e internos dos textos) e **lingüística**
 - Dados externos:
 - cabeçalho que inclui os metadados textuais --- dados bibliográficos comuns, dados de catalogação como tamanho do arquivo, tipo da autoria, a **tipologia textual** e informação sobre a distribuição do *corpus*.
 - Dados internos:
 - anotação de segmentação do texto cru, que envolve:
 - a) marcação da **estrutura geral** – capítulos, parágrafos, títulos e subtítulos, notas de rodapé e elementos gráficos como tabelas e figuras, e
 - b) marcação da **estrutura de subparágrafos** – elementos que são de interesse lingüístico, tais como sentenças, citações, palavras, abreviações e outros elementos relacionados com transcrição (adição, omissão, correção), nomes, referências, datas e ênfases tipográficas do tipo negrito, itálico, sublinhado, etc.
 - Anotação lingüística pode ser em qualquer nível que se queira, isto é, nos níveis morfossintático, sintático, semântico, discursivo, etc...



Dependendo da tarefa/uso ...

- Se um *córpus* é usado para análise sintática (sintagmas nominais),
 - não há necessidade de textos completos
- Se é para o estudo de características do discurso ou para o trabalho terminológico
 - os textos devem ser **completos** o que **nem sempre é necessário** para a **lexicografia**
 - O fato da lexicologia poder trabalhar com trechos de um documento é importante, pois não fere direitos autorais
- Se o *córpus* é para terminologia pode ser menor do que para **lexicografia** que necessita de **grandes *córpus*** para cobrir os vários sentidos/acepções
 - e.g. o vocabulário do inglês é maior do que 1 milhão de palavras e a variedade no uso é grande
 - Por exemplo, a editora Collins tem um *córpus* de 525 milhões de palavras (2005) – o *Bank of English*, que foi lançado em 1991, juntamente com a U. Birmingham.



1. *Desafios no projeto*

- Representatividade é determinada pela variedade de **gêneros/tipos de textos** e como os texto para cada gênero são selecionados
- Um corpus é balanceado se tem um equilíbrio de **gêneros discursivos/tipos de textos** ou de títulos, ou de autores, ou de todos esses itens juntos,
 - desde que as escolhas sejam adequadas à pesquisa que se pretende realizar, demonstrando que os textos foram escolhidos criteriosamente.



1. *Desafios no projeto*

- Como trazer uma variedade de gêneros/tipos textuais em um corpus histórico se a tipologia textual difere das usadas na atualidade?
 - Gênero de textos variam de acordo com a cultura e com o tempo
- Como classificar gênero/tipo de texto corretamente se um mesmo texto, uma carta, por exemplo, cumpria várias funções?
- Como conseguir a variedade e a quantidade se o processo para se ter uma grande quantidade necessária é caríssimo?
 - Escaneamento de fontes impressas e Correção de OCR
 - Digitação de manuscrito
- Textos históricos não estão largamente disponíveis na Web como os textos contemporâneos;
 - o processo de trazê-los para a vida digital preservando todas as características da fonte ainda é muito caro.



1. Solução adotada no DHPB

- Criação de uma Tipologia de Domínios Discursivos e Gêneros Textuais, baseada:
 - em outros projetos de córpus históricos,
 - no livro *Belloto, H.L. Como fazer análise diplomática e análise tipológica de documento de arquivo, 2002.*
 - na experiência com córpus contemporâneos
- Incentivar o preenchimento de domínio e subdomínio; gênero e subgênero, na anotação manual
- Estudar formas de anotação automática deste metadado como uma **pesquisa de mestrado**
 - Usando métodos de aprendizado de máquina supervisionado, como os de Rachel Aires, no seu doutorado, para córpus contemporâneos
 - <http://www.nilc.icmc.usp.br/nilc/projects/linguarudo.html>
 - Usando métodos de aprendizado que agrupam textos com certas características
 - SARDINHA, Tony Berber. Multidimensional analysis. **DELTA**, São Paulo, v. 16, n. 1, 2000 .



Tipologia de Domínios Discursivos

- **8 domínios:**
 - Religioso, Jurídico, Científico, Informativo, Referencial, Instrucional, Técnico Administrativo e/ou Oficial, Literário, Pessoal



7.1. Comunicacional

7.1.1. ato

...

7.1.2. carta

7.1.2.1. carta de apresentação

7.1.2.2. carta régia

7.1.2.3. carta de abrasão de armas de nobreza e fidalguia

7.1.2.4. carta de confirmação

7.1.2.5. carta de conta

7.1.2.6. carta de diligência

7.1.2.7. carta de doação

7.1.2.8. carta de examinação

7.1.2.9. carta de mercê

7.1.2.10. carta de nomeação

7.1.2.11. carta de ofício

7.1.2.12. carta de ordenança

7.1.2.13. carta de prego

7.1.2.14. carta de privilégio

7.1.2.15. carta de propriedade

7.1.2.16. carta de sentença

7.1.2.17. carta oficial

7.1.2.18. carta-relatório

7.1.2.19. carta de alforria

7.1.2.22. carta de sesmaria

7.1.3. circular

7.1.4. declaração

7.1.5. despacho

7.1.6. informação de serviço

7.1.7. memorando

7.1.8. ofício

7.1.9. provisão

7.1.10. requerimento

7.1.11. solicitação

7.2. Descritivo

7.3. Comercial

Técnico
administrativo
e/ou oficial

Subgênero





2. *Desafios na Compilação e Anotação*

- Problemas freqüentes em textos históricos (Rydberg-Cox, 2003; Sanderson, 2006):
 - **Palavras comuns e fins das palavras são abreviados,**
 - usando símbolos tipográficos não comuns - não pertencem ao conjunto ISO 8859-1 (Latin-1) ou estendido
 - **Separação silábica nos fins da linha nem sempre são hifenizadas,**
 - gerando não-palavras
 - **Separação de palavras nem sempre são usadas**
 - a junção gera problemas para a contagem de freqüência
 - **Símbolos tipográficos não comuns**
 - aparecem também em palavras não abreviadas
 - **Grande variação de grafia**
 - até mesmo dentro de um mesmo texto
 - **Critérios de transcrição de manuscritos**
 - variam para os símbolos de inserção/remoção de material [] [?] e < > e para desdobramento de abreviaturas
 - símbolos utilizados tem comportamentos variados com processadores de cópús



.... de Britto

..... de Britto

Auto de inventar[io] que o juis
ordi[nário e dos] or[fãos] antº
Correia da silva mãodou fazer por
falesimento de frº bocado de britto

1650

Nº 44

Muitas
abreviaturas



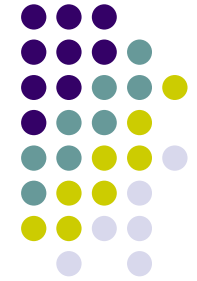
.....
Anno de nasimento de nosso s^{or} jesus xpº de mil e seis sentos e
sincoenta e quatro annos en os trinta dias do mes de marsso da
sobredita era nesta vila de santa anna da parnaiba da cap^{ta} de
são v^{te} estado do brazil Ettª. nesta dita vila nas cazas da morada
que foi de frº bocado de britto que ds ten pelo juis ordinario e dos
orfãos antº correia da silva foi mãodado a min t^{am} e escrivão
fazer este auto p^a por ele eventariar os bês e fazenda que ficou
por morte e falesimêto de frº bocado de britto que d^õ t^e p^a o que
deu juramento dos santos evangelhos a viuva tomazia Ribrª
mulher que foi do dito defunto p^a que sob cargo dele declarasse
e manifestasse todos os bês e fazenda que pesuhia asin moveis
como de rais drº ouro prata joias dividas que se devesen a fazenda
/ e as que a fazenda deve e ela o [pro]met[eu] asin fazer de que
tudo fis este auto en que o dito juis asinou e pela viu[va] não
saber ela o [pro]met[eu] asin fazer de que fis este auto en que o

INVENTÁRIO E TESTAMENTO
DE FRANCISCO BICUDO DE
BRITO - 1654, VILA DE SÃO
PAULO (APENSO O
TESTAMENTO DE TOMÁSIA
RIBEIRO DE ALVARENGA),
SÍLNIÁ NUNES MARTINS,
EDITORA RESPONSÁVEL DA
DIVISÃO DE ARQUIVOS DO
ESTADO DE SÃO PAULO

Anotação de adição
do Editor



Abreviaturas: ambigüidade e variações de grafia das abreviaturas



Expansões de B^o:

bairro

Bartolomeu

bastardo

beco

bento

Bernardo

(...)

Abreviaturas de Janeiro

Jan

Jan.^{ro}

Jan^{ro}

Janr.^o

Jan.^o

Jn^{ro}

Janr^o



Variação da grafia

declaração → fica em juizo dois mil duzentos e cecenta Rs. 2260
Resto do d^o. q emtr<e>gou domingos da
Rocha E christovão pr^a e na entrega della 100
derão menos sem Rs. de q mandou o dito juiz
fazer esta clareza, e o tostão de menos
entregou christovão perr^a. eu joão viegas
escrivão dos orfão o escrevi em os vinte e tres
de abril de mil seis sentos e cetenta e hũ anno -

Variação da grafia

Caracteres não
pertencentes ao
latim básico ou
estendido

fr^a

237

PEDRO CARAÇA, INVENTÁRIO E TESTAMENTO,
1653 - VILA DE SÃO PAULO. APENSO: INVENTÁRIO
E TESTAMENTO DE MARGARIDA RODRIGUES 1634 - VILA DE SÃO PAULO,
SÍLNIA NUNES MARTINS, EDITORA RESPONSÁVEL PELA DIVISÃO DE ARQUIVOS
DO ESTADO DE SÃO PAULO



Formas das Abreviaturas já pré-processadas

- sarg.ºto P.ºe S.ºor S.ºr m.ºto grd.ºe dr.ºo
- q.ºm P.ºe I.ºo V.ºte s.ºor xp.ºo
- @
- 8.bro
- Carv. q. Sr.
- Sñor

o s.ºor jesus xp.ºo (



Palavras hifenizadas

O padre noviço, que acompanhou ao Padre Francisco Veloso, teve mais bom [tempo ?] de experiência nesta peregrinação, porque além da fome, que a caridade fez voluntária e a necessidade forçosa, a praga de mosquitos que neste sítio do Itaqui se padecia, por ainda não estar bem descoberto, era cruel e contínua de noite e de dia. Todo o rôsto e mãos se lhe cobriram ao pobre Padre de tão grandes chagas, feitas das mordeduras, que esteve lá tão gravemente enfermo como pudera de outra qualquer doença. No Padre Veloso, como feito à prova do Brasil, não causou |

Anotação de dúvidas do Editor

CARTA LXVI - AO PADRE PROVINCIAL DO BRASIL
1654, ANTÓNIO VIEIRA , J. LÚCIO D'AZEVEDO (ed.)

reduzirá estes obstinadíssimos ânimos a acomodamento.

A barca que despachou o senhor Embaixador ainda não é partida à causa do vento. De Lisboa não tivemos carta mais que de Mr. Lanier. As novas que V. Ex.^a nos dá, [de ?] em Alentejo se converterem as armas em arados (2), parece

CARTA XVII - AO MARQUÊS DE NIZA 1648 — JANEIRO 12,
ANTÓNIO VIEIRA , J. LÚCIO D'AZEVEDO (ed.)

Critérios adotados na transcrição



Apresentaremos, a seguir, a edição semidiplomática do primeiro fólio de dois diferentes documentos. Para a realização deste tipo de atividade é necessário o estabelecimento de algumas normas, a saber:

1. Respeitar fielmente o texto: grafia (letras e algarismos), linha, fólio, etc;
2. Indicar o número de fólio, à margem direita, fazendo a chamada com asterisco;
3. Numerar o texto linha por linha, indicando a numeração de cinco em cinco, desde a primeira linha do fólio;
4. Separar as palavras unidas e unir as separadas;
5. Desdobrar as **abreviaturas apresentando-as em itálico e negrito**;
6. Utilizar colchetes para as interpolações;
7. Utilizar chaves para as letras e palavras expurgadas;
8. Indicar as rasuras ilegíveis com o auxílio de colchetes e reticências;
9. Expontuar as letras de leitura duvidosa.

(<http://www.filologia.org.br/revista/32/02.htm>)



Critérios adotados na transcrição

Critérios adotados na transcrição

- Respeitar fielmente o texto: grafia (letras e algarismos), linha, fólho, etc.;
- Indicar o número do fólho, à margem direita, fazendo a chamada com asterisco;
- Numerar o texto, linha por linha, indicando a numeração de cinco em cinco, desde a primeira linha do fólho;
- Separar as palavras unidas e unir as separadas;
- Desdobrar as **abreviaturas com o auxílio de parêntesis: ()**;
- Utilizar colchetes para as interpolações: [];
- Utilizar chaves para as letras e palavras expurgadas: { };
- Indicar as rasuras ilegíveis do texto com o auxílio de colchetes e de reticências: [...];
- Expontuar as letras de leitura duvidosa.

(<http://elies.rediris.es/elies13/queiroz.htm>)



Aos dezoito dias do mes de outubro de mil e seis sentos e sesenta Annos nesta
v^a de santa Anna da pernaiba da capitania de são viscente et^e perante o juis
ordinairo dos orfãos ge[org]ge moreira pareseu o capp^m s[a]lvador Bicudo de
mendon[ça] e por elle foi dito que elle devia neste inventairos [três] mil e
duzentos Reis que avia tomado a ganhos o qual dr^o. elle [o]ra vinha a pagar
como de ~~de~~feito logo pagou Requerendo ao dito juis lhe man[da]sse fazer [fl.
35 v.]d[o t]empo [que] teve o dito dr^o. em seu p[oder] que forão Annos eu que
se montarão as ganancias a sentos e doze Reis [qu]e com o prin[ci]pal faz
soma de tres mi] e novesen[tos e] doze [r]eis Requerendo a[o] dito juis lhe
ase[itasse] o [dito] dr^o. e o dezobrigasse a elle e a seu fiador o que visto pelo
dito juis lhe aseitou o [dito] dr^o. e ouve por [d]ezobrigado a elle e a seu fiador
com
declaração que se tirou sem Reis [d]este termo e comtagem de que fiz
este termo em que asinou com o dito juis e eu Ant^o Ro iž de m[att]jos t^m e es [cri]
vão dos orfãos que o escrevi+

Padrões variados de
+ anotação do Editor

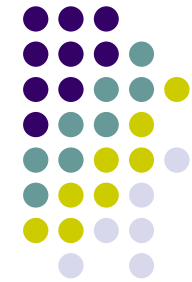
INVENTÁRIO E TESTAMENTO DE GASPAS DIAS PERES (1654),
GASPAS DIAS PERES, SÍLNIA NUNES MARTINS, EDITORA RESPONSÁVEL
DA DIVISÃO DE ARQUIVOS DO ESTADO DE SÃO PAULO

Your search found 22 occurrences

[Click here for a Concordance Report](#)

Occurrences 1-22:

1. A00_0827 (bib:p.0)nselho com hum lutaro ou calujno como de feito fora e lhe contara como ho papa n
2. A00_0827 (bib:p.0) e que viesse ao Rio de Janeiro e como de feito elle se embarquara loguo em huma
3. A00_0827 (bib:p.0)não sera muito que infame outrem como de feito infamou aquelle mancebo dizendo q
4. A00_0059 (bib:p.0)hé contente e lhe praz de dar, e como de feito deu e trespasou, ao dito Collegio
5. A00_1599 (bib:p.0) ajudarião ho melhor que pudesem como de feito fizeram e a outro dia despois di
6. A00_1599 (bib:p.0)rancezes que no dito Rjo estauão como de feito ho dito governador fora e chegand
7. A00_1599 (bib:p.0)m sitio com serquas e balluartes como de feito foi muito grande e boa con casas
8. A00_1599 (bib:p.0)izerem que sua allteza o mandaua como de feito fora em huma armada com muito pou
9. A00_1316 (bib:p.0) era armada inimiga, mas de paz, como de feito era, e estando olhando para ella
10. A00_1316 (bib:p.0)rte: nelle vos poreis em salvo, como de feito succedeu tudo assi. Finalmente, d
11. A00_1316 (bib:p.0)vir á barra inimigos corsarios: como de feito assim aconteceu; mas vendo que a
12. A00_0173 (bib:p.0) dito padre antonio Roiz velho como de feito eizevio E o dito juis lhe da esta
13. A00_0173 (bib:p.0)is os quais os vinha Eizevir he como de feito os Eizevio he de como os eizevio
14. A00_0755 (bib:p.0) dr^o. elle [o]ra vinha a pagar como de feito logo pagou Requerendo ao dito jui
15. A00_1464 (bib:p.0)m nem se lhe consedeçe po | der como de feito se lhe não consede pera | (Fl. 4
16. A00_1463 (bib:p.0) Rez[o]es a | pontadas mandarão como de feito man | dão do primeiro nauio que
17. A00_1466 (bib:p.0) nomeado | que Elle se obrigaua como de feito se obrigou em nome do ditto Bento
18. A00_1466 (bib:p.0)co daguiar que Elle se obrigaua como de feito se obrigou pelo ditto Ben | to da
19. A00_0169 (bib:p.0)elles foi dito, que faziam ora, como de feito fizeram, seu certo, e abundoso, e
20. A00_0169 (bib:p.0) e me apraz de lhe fazer mercê, como de feito por esta presente Carta faça mer
21. A00_1119 (bib:p.0)serviso de Sua Real Magistadi e como de feito tem perdido oito homens brancos,
22. A00_0745 (bib:p.0)por bem e me praz de lhe fazer, como de feito por esta presente carta faça, me



*como de feito X
como de efeito*

Your search found 6 occurrences

[Click here for a Concordance Report](#)

Occurrences 1-6:

1. A00_0171 (bib:p.0)esenta reis os quais apresentava como de efeito a presente e o dito juis o aseit
2. A00_0171 (bib:p.0)o avia pagos o que a gan[hos]... como de efeito logo pagou em dr^o. de contado {
3. A00_0171 (bib:p.0) a ganhos os quais vinha a pagar como de efeito em dr^o. de contado requerendo a
4. A00_0171 (bib:p.0)nta e sinco mil as quais a pagar como de efeito de contado e a seu fiador o que
5. A00_0755 (bib:p.0)dos orfãos o qual, dr^o. trazia, como de efeito logo trouxe e entregou ao ditto
6. A00_0755 (bib:p.0)ta Reis q[ue] ora, vinha a pagar como de efeito logo pagou em dr^o. de con[tado]

**Supondo < > sendo
eliminação**

Símbolos da transcrição atrapalham a busca



- Por exemplo, caracteres "[]" (colchetes) e "< >" (colchetes angulares)
- Exemplos: "<e>feito", "s[a]lvador"
- Semântica do editor não é a mesma das ferramentas
- No Philologic "[]" indica vários elementos de um conjunto (expressão regular).
- No Unitex "<e>" indica cadeia vazia (expressão regular).
- **Buscar:**
 - No Unitex: "s[a]lvador" e "\<e\>feito"
 - No Pilologic: "s a lvador" ("<e>feito" não pode ser buscado no Philologic)



ções que lhe insinamos, e nom parece honesto estarem nuas
235 entre os christãos na igreja, e quando as insinamos. E disto
peço ao P.^e M. João²¹ tome cuidado, por elle ser parte na
conversão destes gentios, e nom fique senhora nem pessoa
a que nom importune [5r] para cousa tam sancta, e a isto se
avião de aplicar todas as restituções que lá se ouvessem
240 de fazer, e isto agora soamente no começo que elles farão
algodões para se vestirem do diante.

Notas de
Rodapé

14. Os Irmãos todos estão de saude e fazem o officio a
que forão enviados: somente Antonio Pirez se acha mal das
pernas, que lhe arebentarão depois das maleitas²² que teve,
245 e nom acaba de ser bem são.

Variações
de grafias

Leonardo Nunez mandei aos Ilheos, huma povoação
daqui perto, onde dá muito exemplo de si e faz muito fruto,
e todos se spantão de sua vida e doctrina. Foi com elle
Diogo Jácome, que faz muito fruto em insinar os moços e
250 escravos.

15. Agora pouco há vierão aqui a consultar-me algu-
mas duvidas, e esteverão aqui por dia do Anjo²³, onde

Mais variações de grafia complicando a contagem da frequência de palavras do corpus ...

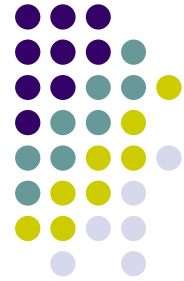


que lhe **insinamos**, e **nom** parece honesto estarem nuas entre os **christãos** na igreja, e quando as **insinamos**. E disto peço ao P.^o e M. João tome cuidado, por **elle** ser parte na conversão destes gentios, e **nom** fique senhora nem pessoa a que **nom** importune [5r] para **cousa tam sancta**; e a isto se **avião** de **applicar** todas as restituições que lá se **ouvessem** de fazer, e isto agora **soamente** no começo que **elles** farão algodões para se vestirem ao diante.

14. Os Irmãos todos estão de **saude** e fazem o **officio** a que forão enviados: somente Antonio Pirez se acha mal das pernas, que lhe **arebentarão** depois das maleitas que teve, e **nom** acaba de ser bem são. Leonardo Nunez mandei aos Ilheos, **huma** povoação daqui perto, onde dá muito exemplo de si e faz muito **fruito**, e todos se **spantão** de sua vida e **doctrina**. Foi com **elle** Diogo Jácome, que faz muito **fruito** em **insinar** os moços e escravos.

CARTA DO P. MANUEL DA NÓBREGA AO
P. SIMÃO RODRIGUES, BAÍA 9 DE AGOSTO 1549, SERAFIM LEITE S. J (ed.)

daCamera: 58 ocorrências; daCamara: 25; complicam a contagem da frequência de palavras



[home](#) [the ARTFL project](#) [download](#) [documentation](#) [sample databases](#)

Your search found 58 occurrences

[Click here for a Concordance Report](#)

Occurrences 1-58:

1. A00_1380 (bib:p.0) s nesta Cidade do Salvador e Casas da Camera estando presentes os Offeciaes da Camera
2. A00_1380 (bib:p.0) a estando presentes os Offeciaes da Camera mandaram fazer este assento em que asenta
3. A00_1379 (bib:p.0) r por seu Despacho que os Offeciaes da Camera desta Cidade aquem he notorio tudo o que
4. A00_1379 (bib:p.0) - E Recebera Merce - Os Offeciaes da Camera - vejam esta Peticao, edem Casas ao Supli
5. A00_1378 (bib:p.0) nesta Cidade do Salvador e Casas da Camera estando ahi os Officiaes da Camera manda
6. A00_1378 (bib:p.0) zes ficando agosto dos Offeciaes da Camera para o que dara fianca assim para adit
7. A00_1378 (bib:p.0) di coens, e asinou com os Offeciaes da Camera, e Porteiro e Eu -- Ruy de Carvalho Pinhe
8. A00_1381 (bib:p.0) nesta Cidade do Salvador, e Casas da Camera, estando ahi os Offeciaes della se asento
9. A00_1381 (bib:p.0) ue asinou com os dittos Offeciaes da Camera - Ruy de Carvalho Pinheiro o Escrevy
10. A00_1381 (bib:p.0) lararam mais os dittos Offeciaes da Camera, que asim setiraram do Monte Mor todos o
11. A00_1377 (bib:p.0) Camera estando ahi os Officiaes da Camera asaber os Juizes Francisco de Barbuda, e A
12. A00_1377 (bib:p.0) fica dito pelo que os Officiaes da Camera em nome do Povo se obrigarão a comprar ad
13. A00_1376 (bib:p.0) nesta Cidade do Salvador e Casas da Camera estando ahi o Juiz Frãcisco de Barbuda,
14. A00_1375 (bib:p.0) nesta Cidade do Salvador, e Casas da Camera estando em Vereação os Officiaes da Cam
15. A00_1383 (bib:p.0) nesta Cidade do Salvador e Casas da Camera, estando ahi os Juizes Cosme de Sá Peixot
16. A00_1383 (bib:p.0) adar da propria que esta na area da Camera que mereporto hoje sete dias domez de Abr
17. A00_1384 (bib:p.0) r Pedro da Silva com os Offeciaes da Camera sobre odinheiro que setirou da Entrada do
18. A00_1384 (bib:p.0) ena de Vilha San ti eos Offeciaes da Camera que servem este presente anno asaber Cos
19. A00_1384 (bib:p.0) de com os sobredittos Offeciaes da Camera se repartiram namaneira seguinte quatro c
20. A00_1384 (bib:p.0) to Senhor Governador, e Offeciaes da Camera applicados para segastarem nos quarteis d
21. A00_1384 (bib:p.0) o Senhor Governador, e Offeciaes da Camera que setornassem por emprestimo para a faz
22. A00_1384 (bib:p.0) uy de Carvalho Pinheiro Escrivam da Camera desta Cidade do Salvador fiz trasladar da
23. A00_1385 (bib:p.0) nesta Cidade do Salvador e Casas da Camera, estando ahi os Offeciaes della, e junta
24. A00_1385 (bib:p.0) ovo, e asinaram com os Offeciaes da Camera, e Eu Ruy de Carvalho Pinheiro o Escrevi
25. A00_1386 (bib:p.0) esta Cidade do Salvador, e Casas da Camera digo do Salvador Bahia de todos os Santos,
26. A00_1386 (bib:p.0) do Estado do Brazil, e Offeciaes da Camera, eos Cidadões, e Pessoas desta Cidade aba

Junção de palavras - Eque: 79 ocorrências complicam a contagem da frequência de palavras



Your search found 79 occurrences

[Click here for a Concordance Report](#)

Occurrences 1-79:

1. A00_1372 (bib:p.0)is desse, na forma da Ordenação, eque aobrigação de acodir poristo, carregava so
2. A00_1374 (bib:p.0)aver tempo de semandar apregoar, eque as ditas penas seexecutarão logo conforme o
3. A00_1376 (bib:p.0)orcanada devinho heram acabados, eque o GovernadorPedro daSilva npstinha represent
4. A00_1382 (bib:p.0)endo-lhe praticas que heram meus eque VossaSenhoria mosdéra poreu a situar aquell
5. A00_1382 (bib:p.0)de Tapiragua com osdepatigipeba, eque não haja falta: mandandome VossaSenhoria as
6. A00_1383 (bib:p.0)adacanada devinho eram passados, eque oGovernador Geral deste Estado Pedro da Sylv
7. A00_1393 (bib:p.0)rem carnes para oditto Sustento, eque seodito dinheiro senam applicasse para isso n
8. A00_1384 (bib:p.0)por estar muito impossibilitada, eque secarregasem em Receita aoThezoureiro Geral
9. A00_1386 (bib:p.0)oproveito, foce geral o encargo, eque suposto que aFazendaReal estava emtanto aper
10. A00_1386 (bib:p.0)fazerse cada anno esta despeza, eque para sesaber o quanto será necessario para
11. A00_1389 (bib:p.0)restimo de quinze mil cruzados, eque deprezente, epor ora os emprestasem oshomen'
12. A00_1389 (bib:p.0)as que fizeram odito imprestimo eque aditta repartiçam sefazia por duas pessoas
13. A00_1399 (bib:p.0)mada Real que esta neste porto, eque este povo em tempo menos apertado doque hoje
14. A00_1398 (bib:p.0)ue correrá com assentoz crenas, eque dará contas com clarezadetudo quanto nella
15. A00_1395 (bib:p.0)coidado plantasemse mantimentos eque por senam haver feito havia deprezente tam g
16. A00_1395 (bib:p.0) epassar melhor afalta prezente,eque fazendo elle Conde General consideraram em o
17. A00_1395 (bib:p.0)detodas as Capitaniaz do Norte, eque conforme apossibilidade decada hum os obriga
18. A00_1395 (bib:p.0) aquantidade que lhes tocassem, eque seriamaisconveniente que cada qual deseos ne
19. A00_1395 (bib:p.0)haveriam as Ordens necessarias, eque selhe porião pennas aos que nam plantassem
20. A00_1395 (bib:p.0) de angolla, eduzentoz cruzados eque assim ficava bastante mente previnido emquan
21. A00_1395 (bib:p.0)ernambuco tinham muitos negros, eque ostraziam alugadoz adiferentes fabricas, equ
22. A00_1395 (bib:p.0) alugadoz adiferentes fabricas, eque convinha obrigarallos aq. plantassem, seassent
23. A00_1395 (bib:p.0)ametade das fabricas que tinham eque assim lheficasem aoutra ametade para os alug
24. A00_1394 (bib:p.0) rte que o inimigo tem occupado, eque podia aodiante haver mais faltas demantiment
25. A00_1397 (bib:p.0) Praça ebem dos moradores dela, eque aconservaçam detudo dependia de haver armad
26. A00_1397 (bib:p.0) icaria este danno inremediavel, eque para Sustentar adita armada a Fazenda Real d
27. A00_1397 (bib:p.0)do a Recuperaçam de Pernambuco, eque emfaltando Sesçava esta obrigaçam como que
28. A00_1403 (bib:p.0) tinha concedido para as crennas eque isto sefizesse por Repartição igual sem af
29. A00_1403 (bib:p.0)sta he averdadeira fortificação eque se ouvira quando digo eque se os ouvera quand
30. A00_1403 (bib:p.0)scara, efizera osdannos quefez, eque lhesfosse defabrica, e apresto dos Galiatoz
31. A00_1403 (bib:p.0)ia selhe buscaram, ese lhederam,eque omesmo sopodera fazer aguerra seos dittoz Ga
32. A00_1403 (bib:p.0)ão bem oseu effeito, eobrigaçam eque pelladitta razão não deviam ser obrigaradoz

Coordenar o trabalho de uma grande equipe que

- **Faz a seleção dos textos que comporão o corpus**
- **Escaneia e corrige erros de OCR**
- **Preenche cabeçalho com vários metadados**
- **Trata hifenização**
- **Pré-processa os textos para serem usados por processadores de corpus**
- **Adapta processadores de corpus para tratar da escalabilidade e funcionalidades adequadas à tarefa**
- **Anota fenômenos lingüísticos com padrões internacionais para que o corpus possa ser útil para outros projetos**

....criar um corpus de textos históricos é uma empreitada cara e demorada, portanto este tipo de corpus deve ser reusado por outros grupos de pesquisa e/ou outros projetos





2. *Soluções adotadas no DHPB*

- **Anotação dos metadados e dos textos. Uso de padrões internacionais: TEI (cabeçalho, notas, junção).**
 - Notas dos editores devem ser tratadas, pois não fazem parte do texto
- **Codificação de caracteres que caíram em desuso. Uso do Unicode e padronização as escolhas dos códigos**
- **Abreviaturas. Uso de um dicionário de abreviaturas no formato DELA do Unix para pesquisa quando há dúvidas do significado**
- **Variação de grafia. Criação de um sistema (SIACONF) para agrupar grafias e codificação delas num dicionário de variações de grafia no formato DELA do Unix: ajuda a preencher campos do verbete**
- **Junção das palavras. Identificação delas com a ajuda de um filtro do corpus com um dicionário contemporâneo do PB e anotação da separação delas com padrões internacionais (TEI)**

Padrões Internacionais de Anotação e Codificação



- Como o custo de se criar *córpus* anotados é muito alto
 - tanto em termos financeiros como na demanda de trabalho especializado,
- pesquisadores amortizam estes custos reusando estes recursos
- Este alto custo contribui para o desenvolvimento de padrões de codificação e anotação
 - para recursos de língua, que permitem o seu intercâmbio
- Exemplos de padrão de anotação:
 - **TEI** – mais adaptado para *córpus* históricos e
 - **XCES** – mais adaptado para criação de *córpus* para PLN
- Padrão de codificação de caracteres: Unicode
- Vantagens de se usar estes padrões internacionais:
 - Facilita o intercâmbio de dados, reuso e extensibilidade
 - Evita o desenvolvimento de software, pois podemos usar ferramentas já desenvolvidas que os atendem



Para *córpus* históricos ...

- **Unicode é fundamental, pois permite a representação de caracteres que caíram em desuso**
- **Como o conjunto de símbolos é muito extenso, precisamos delimitar um conjunto.**
- **Por exemplo, há vários códigos para o til diacrítico, escolhemos o 0303**
- **o pode ser codificado como:**
 - grau (00B0), “zero” sobrescrito (2070), “o” sobrescrito (00BA), anel (02DA), entre outros,
 - escolhemos “o” sobrescrito (00BA),
- **a foi codificado como “a” sobrescrito (00AA)**

Symbol	Description	Unicode	Sample
^	combining circumflex accent	0302	quarÿ (*)
~	combining tilde	0303	corñandante (commander)
¯	combining macron	0304	cacaō (cocoa beans)
¨	combining diaeresis	0308	muÿ (much, many)
◌̆	combining hook above	0309	sõmente (only)
◌̇	Combining ring above	030A	Å (abbreviation of Afonso)
◌̈	Combining comma above	0313	tinhab
Æ	Latin capital letter AE	00C6	Æthyopia (**)
æ	Latin small letter ae	00E6	grati (**)
Œ	Latin small ligature oe	0153	Cœteris (**)
§	section sign	00A7	§ (denotes paragraph mark)
ƒ	turned capital f	2132	
ſ	Latin small letter long s	017f	Deſcobrio (find)
ƒ	Latin small letter f with hook	0192	
Ʒ	Latin small letter turned e	01DD	
Ɔ	Latin small letter turned a	0250	

(*) Indian name (**) Latin name

Table 5: Characters found in historical texts.

**Escolhas do Projeto DHPB
para diacríticos e outros
símbolos**

Tratamento do Sobrescrito em abreviaturas



(...) apartida de belem como vosa alteza sabe foy **seg^a** feira ix demarço. e sabado xiiij do dito mes amtre as biiij e ix oras nos achamos amtre as canareas mais perto da gram canarea e aly amdamos todo aquele dia em calma avista delas obra de tres ou quatro legoas. e domingo xxij do dito mes aas x oras pouco mais ou menos ouuemos vista dasilhas do cabo verde. s. dajlha de sã njcolaa**o seg.^o** dito de **p^o** escolar piloto. e anoute segujmte **aaseg^{da}** feira lhe (...)



(...) apartida de belem como vosa alteza sabe foy **seg^a** feira ix demarço. e sabado xiiij do dito mes amtre as biiij e ix oras nos achamos amtre as canareas mais perto da gram canarea e aly amdamos todo aquele dia em calma avista delas obra de tres ou quatro legoas. e domingo xxij do dito mes aas x oras pouco mais ou menos ouuemos vista dasilhas do cabo verde. s. dajlha de sã njcolaa**o seg.^o** dito de **p^o** escolar piloto. e anoute segujmte **aaseg^{da}** feira lhe (...)



Notas nos textos históricos

7. - BAÍA 9 DE AGOSTO DE 1549 127

convertidos, onde estaremos Vicente Rodriguez e eu, e hum soldado¹⁹ 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 210 hi para isto. Os officiaes que cá estão tem muito que fazer, e que o nom tenham, estão com grande saudade do Reyno, porque deixão lá suas mulheres 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. 215 Portanto me parece que avião de vir de lá, e, se possivel fosse, com suas mulheres e filhos, e alguns que fação taipas e carpinteiros. Cá está hum Mestre para as obras, que hé hum sobrinho²⁰ de Luis Diaz, mestre das obras d'El-Rey, ho qual veo con trinta mil reis de partido.

(...)

19 Simão Gonçalves. LEITE I 573.

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

Notas anotadas em TEI



<p> {7. - BAÍA 9 DE AGOSTO DE 1549 127 - A00_0002.txt,.N} </p>

<p> 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. </p>

<p> O Padre Antonio Pirez e o P.^o 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 tenham, 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. </p>

(...)



Anotação de Cabeçalho TEI

É dividido em 4 elementos principais.

<fileDesc>

Contém uma completa descrição bibliográfica do texto eletrônico. **Obrigatório**

<encodingDesc>

Contém informações sobre a maneira como o texto foi codificado. **Recomendado.**

<profileDesc>

Contém informações sobre vários aspectos do texto (língua usada, classificação do texto segundo a sua tipologia, os participantes de um texto falado e sua situação, anotações, etc.). **Opcional.**

<revisionDesc>

Resume o histórico de revisão (cabeçalho, segmentação e lingüística) de um texto. **Opcional.**

Cabeçalho TEI



```
<?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 DE PERO VAZ DE CAMINHA </title>
      <author>
        <name> PERO VAZ DE CAMINHA </name>
        <date> 01 DE MAIO DE 1500 </date>
      </author>
      <respStmt>
        <resp>Arquivo preparado por</resp>
        <name>varios pesquisadores do Projeto DHPB</name>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <distributor>Projeto do Dicionario Historico do Portugues do Brasil (DHPB), UNESP, Araraquara</distributor>
      <date> 2006 </date>
    </publicationStmt>
    <sourceDesc>
      <biblStruct>
        <monogr>
          <title> CARTA DE PERO VAZ DE CAMINHA </title>
          <author> PERO VAZ DE CAMINHA </author>
          <imprint>
            <pubDate> 1964 </pubDate>
          </imprint>
        </monogr>
      </biblStruct>
    </sourceDesc>
  </fileDesc>
</teiHeader>
<text>
```

```
<teiHeader>
  <fileDesc>
    <titleStmt> ... </titleStmt>
    <publicationStmt> ... </publicationStmt>
    <sourceDesc> ... </sourceDesc>
  </fileDesc>
</teiHeader>
```

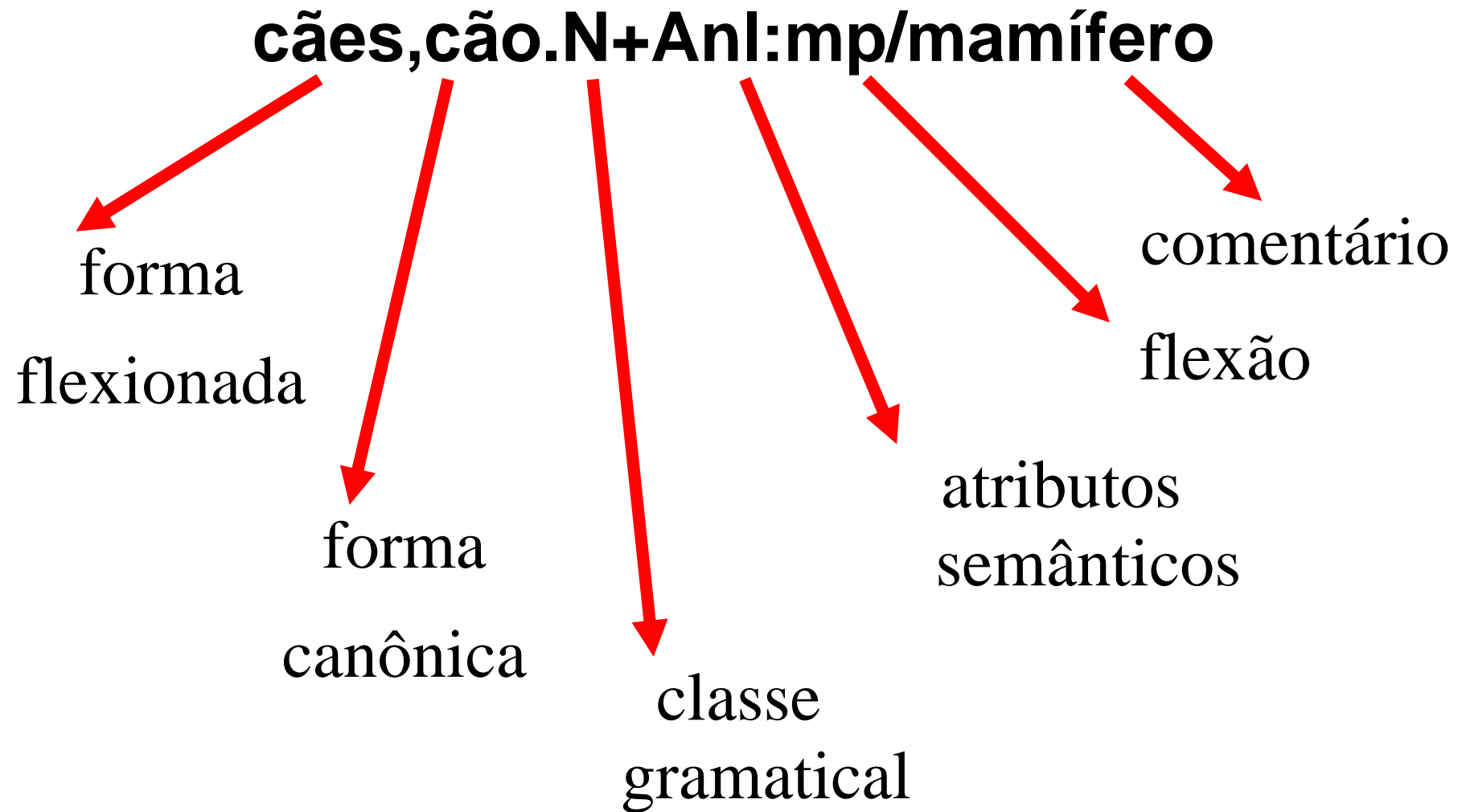
**Cabeçalho
Mínimo**

Limpeza e anotação



- **Protew-lite e Protej – criados por um **mestrado do ICMC****
 - Tratamento de sobrescrito e de formatação em geral
 - Conversão da ficha catalográfica para TEI-Lite
 - Anotação de notas de rodapé, numeração de páginas, parágrafos
 - Entre outros

Léxicos no formato DELA



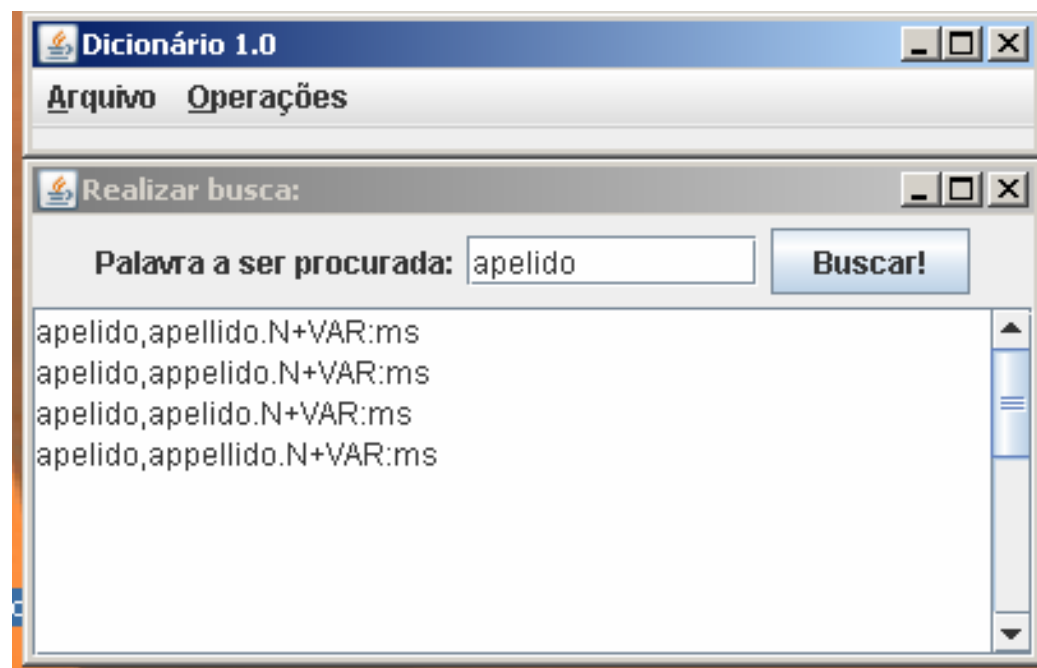
Dic DELA para variantes



apellidos,apelidos.N+VAR:ms/50.0%
apelidos,apelidos.N+VAR:ms/36.36%
apellidos,apelidos.N+VAR:ms/9.09%
apellidos,apelidos.N+VAR:ms/4.54%

Alternativa: apellidos,apelidos.N+VAR+apelido:ms/50.0%

- Todas as entradas são nomes (N) e estão no masculino singular (ms) porque o processo foi automático
- Para gerar o Dicionário de variantes, invertemos os 2 primeiros campos para facilitar a busca, que é feita pelo primeiro campo
- O comentário se perde no formato binário – sugestão discretizar a frequência e colocá-la como atributo semântico
- Variantes vieram do SIACONF



**Pesquisa sendo desenvolvida:
tratamento de verbos e
suas variantes**



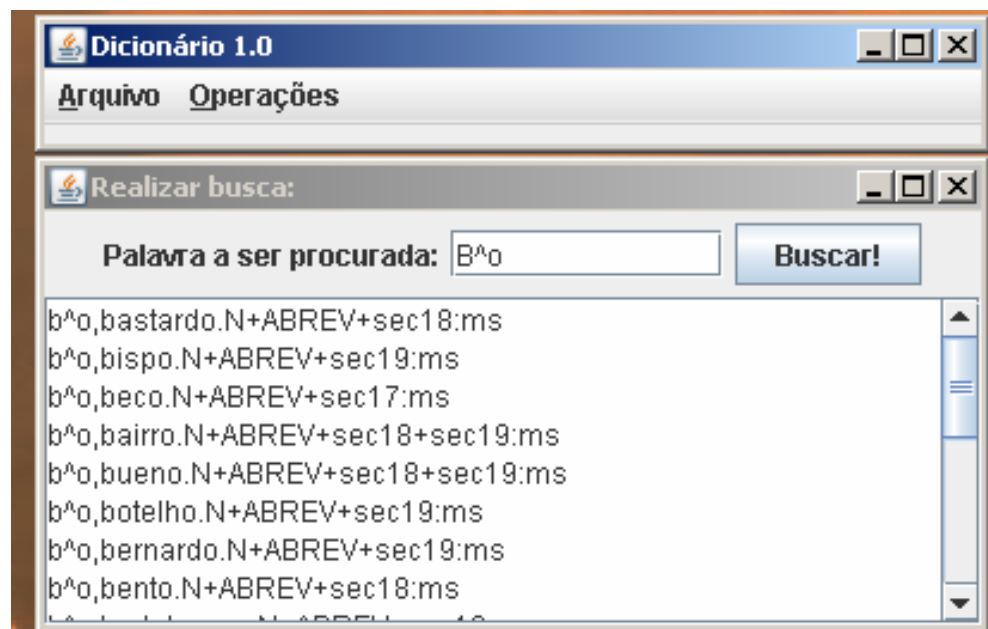
Dic DELA para abreviaturas

a^al,auxiliar.A+ABREV+sec18:fs
a^al,auxiliar.A+ABREV+sec18:ms
a^al,auxiliar.N+ABREV+sec18:fs
a^al,auxiliar.N+ABREV+sec18:ms
a^al,auxiliar.V+ABREV+sec18:U1s
a^al,auxiliar.V+ABREV+sec18:U3s
a^al,auxiliar.V+ABREV+sec18:W1s
a^al,auxiliar.V+ABREV+sec18:W3s

- Tratamos a ambigüidade categorial para algumas letras
- Abreviaturas vieram de:

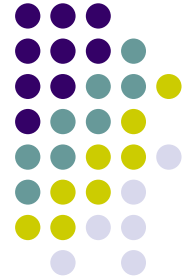
- **FLEXOR**, Maria H. *Abreviaturas, Manuscritos do século XVI ao XIX*. Editora Unesp – secretaria do Estado da Cultura – Arquivo do Estado de São Paulo, 1991.

- outras fontes



Pesquisa: completar a anotação Morfossintática; realizar anotação de Entidades Nomeadas (EN) para que o dicionário seja uma fonte num sistema de extração de EN

SIACONF (Sistema de Suporte para a Contagem de Frequência)



- **Disponível livremente:**
 - <http://moodle.icmc.usp.br/dhpb/siaconf.tar.gz>
- **43 regras de transformação aplicadas em 4.9 milhões de palavras**
 - **12.189** agrupamentos
 - **27.199** variantes
- Baseado nos trabalhos:

Tais A. Menegatti e Helena Britto. “Regras Lingüísticas para Tratamento Computacional da Variação de Grafia e Abreviaturas do Corpus Tycho Brahe”. *Relatório de Iniciação Científica*. UNICAMP (2002)

Alexandre Hirohashi e Marcelo Finger. “Aprendizado de regras de substituição para normalização de textos históricos”. *Dissertações do Instituto de Matemática e Estatística*. Universidade de São Paulo (2005)

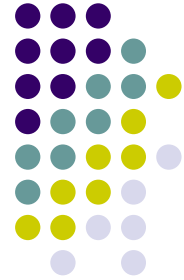
Exemplos de Regras Criadas



- **Six classes of rules created:**
- **1. Rules to deal with spellings that fell in disuse (4 rules)**
 - Example: all "ph" are replaced to "f", because in "ph" is no longer used
 - **phármacia -> fármacia**



- **2. Rules to deal with double consonants (13 rules)**
 - Example: **ffoy** -> **foi**, **edittou** -> **editou**
- **3. rules according orthographic norm (6 rules)**
 - Example: "n" must be replaced by "m" before "b" or "p"
 - **tenpo** -> **tempo**



- **4. Rules based on frequency analysis (14 rules)**
 - Example: replace "ch" by "x"
 - **Cham -> xam**
- **5. Rules used in Tycho Brahe (5 rules)**
 - Example: "z" by "s" in the infix "preciz"
 - **preciza -> precisa**



- **6. Lexicalised rules (1 rule): specific rules to cover spellings which are not grouped by general rules**
 - Example: replace "o" by "u" to forms ending in "deos"
 - deos -> deus, judeos -> judeus

Pesquisa: criar mais regras a partir dos relatórios de apoio do sistema



Exemplos de agrupamentos

<code>apelido (90)</code> <code> appellido (48)</code> <code> apelido (30)</code> <code> appelido (7)</code> <code> apellido (5)</code>	<code>nam (37,100)</code> <code> não (33,684)</code> <code> naõ (2,652)</code> <code> nam (439)</code> <code> nao (325)</code>
<code>mais (23053)</code> <code> mais (22,918)</code> <code> majs (67)</code> <code> maes (38)</code> <code> mays (30)</code>	<code>vila (5,218)</code> <code> villa (4,073)</code> <code> vila (1,113)</code> <code> vyla (13)</code> <code> vjlla (9)</code> <code> vylla (9)</code> <code> vjla (1)</code>

Tratando a junção de palavras



- 1) Busca delas na lista de palavras desconhecidas do Unitex**
- 2) Checagem via concordanceador**
- 3) Anotação com etiquetas TEI**
- 4) Troca automática no corpus da junção pela separação**

Lista de palavras desconhecidas dos dicionários aplicados no Unitex



The screenshot shows the Unitex 1.2 interface with the current language set to Historical Portuguese (Brazil). The main text window displays a document with the following content:

{corpus1.txt,.N}
{A00_0001.txt,.N}
{108 P. MANUEL DA NÓBREGA - P. SIMÃO RODRIGUES - A00_0001.txt,.N}
5
DO P. MANUEL DA NÓBREGA
AO P. SIMÃO RODRIGUES, LISBOA.
BAÍA [10 ? DE ABRIL] 1549
{5.-BAÍA 19 DE ABRIL DE 1549 109 - A00_0001.txt,.N}
A graça e amor de N. Senhor Jesu Christo seja sempre em nosso favor e ajuda. Amen.
1. Somente darey conta a V. R. de nossa chegada a esta terra, e do que nella fizemos e esperamos fazer em ho Senhor Nosso, deixando os fervores de nossa prospera viagem aos Irmãos que mais em particular a notaram
Chegamos a esta Baya a 29 dias do mes de Março de 1549. Andamos na viagem oito somanas. Achamos a terra de paz e quarenta ou cinquenta moradores na povoação que antes era. Receberam-nos com grande alegria; e achamos
{110 P. MANUEL DA NÓBREGA - P. SIMÃO RODRIGUES - A00_0001.txt,.N}
huma maneira de igreja , junto da qual logo nos apousetamos hos Padres e Irmãos em humas casas a par della, que nam foy pouca consolação para nós, para dizermos missas e confessarmos; e nisso nos ocupamos agora. Confessa-se toda haa gente da armada, digo a que vinha nos outros navios, porque os nossos determinamos de hos confessar na nao.
2. Ho primeiro domingo que dissemos missa foy a 4^a. domingo da Quadragessima . Disse eu missa cedo e todos os Padres e Irmãos confirmamos os votos que tinhamos feitos e outros de novo com muita devação e conhecimento

The 'Token list' window shows the following data:

Count	Token
2281145	
181992	,
101935	de
96850	e
90186	.
87997	que
62724	a
53374	o
39258	se

The 'Word Lists in C:\Documents and Settings...' window shows two columns of words:

DLF: 49124 simple-word l...	ERR: 51850 unknown sim...
@,anos.N+ABREV:ms	entrancia
a,.ABREV:ms	entranhámos
a,.N:ms	entranhavão
a,.PREP	entranhavel
A,alteza.N+ABREV:fs	entranhavelmente
A,alvará.N+ABREV:ms	entranhávelmente
A,Amaro.Npr+ABREV:r	entranhos
A,Ana.Npr+ABREV:fs	Entrão
	entraõ
	entrão
	entrará
	entrára
	entrarão
	ENTRÁRÃO
	Entrarão
	entrarão
	entrários
	Entraron
	entrase

Palavras Desconhecidas





Parte da Lista de junção

387.+asdispezas	<choice> <sic> asdispezas </sic> <corr> as dispezas </corr> </choice>
388.+aSegurança	<choice> <sic> aSegurança </sic> <corr> a Segurança </corr> </choice>
389.+asemana	<choice> <sic> asemana </sic> <corr> a semana </corr> </choice>
390.+aseu	<choice> <sic> aseu </sic> <corr> a seu </corr> </choice>
391.+aseupoder	<choice> <sic> aseupoder </sic> <corr> a seu poder </corr> </choice>
392.+aseus	<choice> <sic> aseus </sic> <corr> a seus </corr> </choice>
393.+aSexta	<choice> <sic> aSexta </sic> <corr> a Sexta </corr> </choice>
394.+asfolhetas	<choice> <sic> asfolhetas </sic> <corr> as folhetas </corr> </choice>
395.+asfontes	<choice> <sic> asfontes </sic> <corr> as fontes </corr> </choice>
396.+asforças	<choice> <sic> asforças </sic> <corr> as forças </corr> </choice>
397.+asidade	<choice> <sic> asidade </sic> <corr> a sidade </corr> </choice>
398.+asinadapor	<choice> <sic> asinadapor </sic> <corr> asinada por </corr> </choice>
399.+asmais	<choice> <sic> asmais </sic> <corr> as mais </corr> </choice>
400.+asmesmas	<choice> <sic> asmesmas </sic> <corr> as mesmas </corr> </choice>
401.+asnecessidades	<choice> <sic> asnecessidades </sic> <corr> as necessidades </corr> </choice>
402.+asobras	<choice> <sic> asobras </sic> <corr> as obras </corr> </choice>
403.+asobrigaçõens	<choice> <sic> asobrigaçõens </sic> <corr> as obrigaçõens </corr> </choice>
404.+asomma	<choice> <sic> asomma </sic> <corr> a somma </corr> </choice>
405.+asOrdens	<choice> <sic> asOrdens </sic> <corr> as Ordens </corr> </choice>
406.+aspartes	<choice> <sic> aspartes </sic> <corr> as partes </corr> </choice>

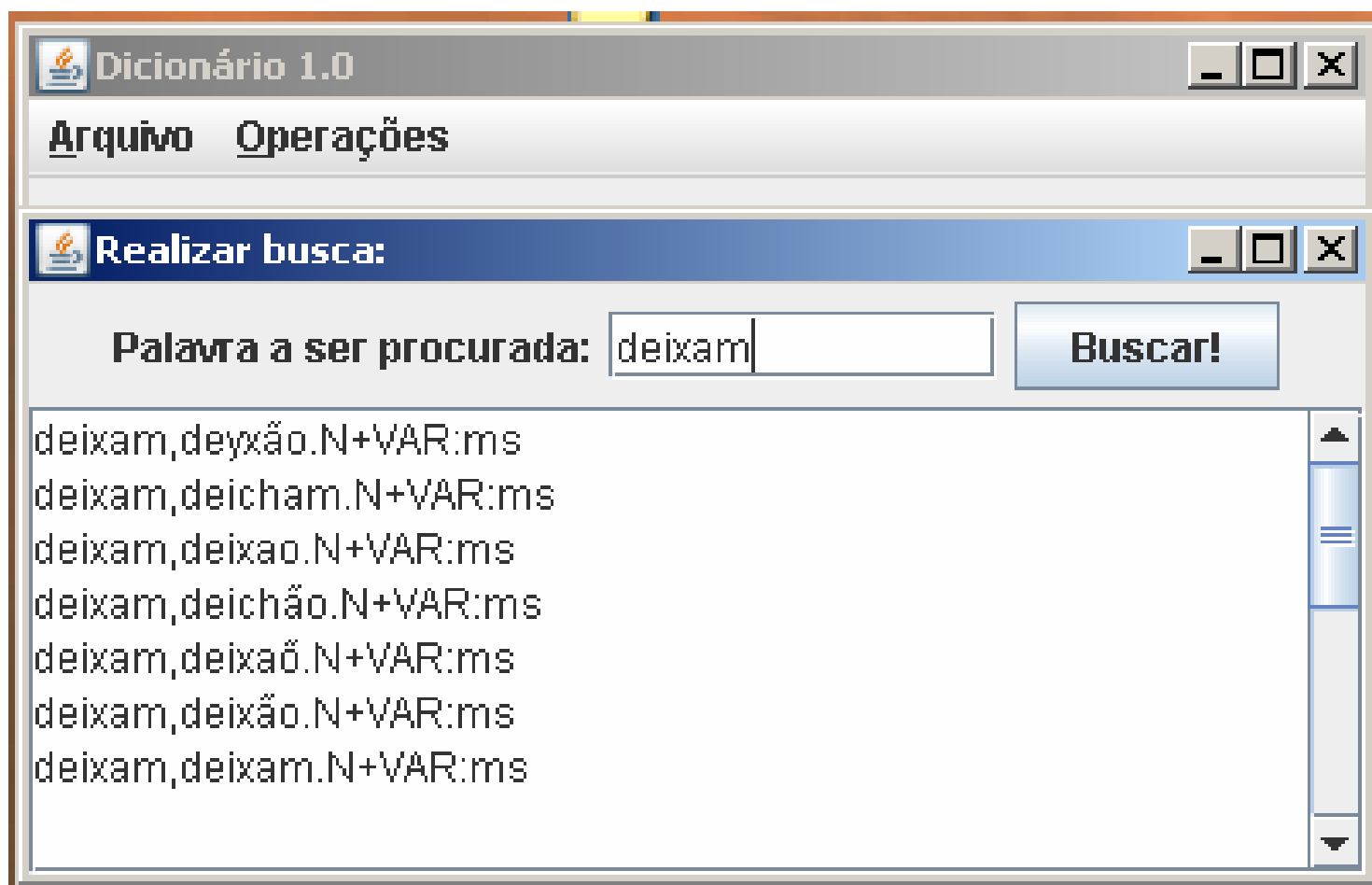
Busca de Variantes



- Na literatura encontramos:
 - **Sistemas baseados em regras como o SIACONF**
 - **Sistemas baseados em distância de edição e outras medidas estatísticas como o AGREP no Philologic (independentes de língua)**
 - **Sistemas híbridos como VARD (inglês) e RSNSR (alemão)**
 - **Sistemas que aprendem a criar regras como o normalizador do Tycho Brahe**
- **No DHPB usamos o Philologic, Dicionário com as variantes trazidas do SIACONF e também as palavras desconhecidas do PB contemporâneo no Unitex.**



Exemplo de uso no Dicionário: deixam



Exemplo de uso na busca por similaridade: deixam



ex futuris neq; ad illo
ad Neapolitanum ea; *Facilia.*
rtim fasid oblietio

PhiloLOGIC

Welcome to PhiloLogic

[home](#) | [the ARTFL project](#) | [download](#) | [documentation](#) | [sample databases](#)

Found 12 matches, shown with frequencies in entire database.

Select words to search in the entire database. Select output options and bibliographic criteria below.

or

98	<input type="checkbox"/>	deitam
378	<input type="checkbox"/>	deixa
13	<input type="checkbox"/>	deixae
4	<input type="checkbox"/>	deixai
9	<input type="checkbox"/>	deixal
278	<input type="checkbox"/>	deixam
2	<input type="checkbox"/>	deixamo
661	<input type="checkbox"/>	deixar
91	<input type="checkbox"/>	deixar
8	<input type="checkbox"/>	deixas
59	<input type="checkbox"/>	deixem
1	<input type="checkbox"/>	meixam

Busca pelo radical no Unitex << ^deix >>



Unitex 2.0beta (July 19, 2007) - current language is Historical Portuguese (Brazil)

Text DELA FSGraph Lexicon-Grammar Edit File Edition Windows Info

Concordance: C:\Documents and Settings\sandra.SANDRA1\Desktop\Trabalho_Unitex\Historical Portuguese (Brazil)\Corpus\corpus_milenio_snt\concord.html

tenha cuidado do gentio que de todo não [deixe](#) os brancos, e assy os visitão de tarde em tarde,
: razão é que também ele, Nóbrega, «não [deixe](#) cousa de consolação ou desconsolação de que lhe n
r quem tenha semelhantes imaginações as [deixe](#) e tenha por falsas e venha ajudar seus Charissimo
o menos escandalizado dos christãos, me [deixei](#) ficar e V. M. se tornou em paz. 10. Nesta Capita
da Prata. Vivendo eu com este desejo, o [deixei](#) de pôr por obra por não ter quem mandar e alguma
uaresma & satisfazer aa devaçoão sua não [deixei](#) de pregar aquelle dia, ainda que 155 aa tarde ou
ivem da mesma maneira; mas com tudo não [deixei](#) o Advento passado e a
à cidade. Mas eu declinando o foro não [deixei](#) de o emccarrar, nem Simão (o
a Prata. Vivendo eu com este desejo, ho [deixei](#) de pôr por obra por não ter
mas que eu trazia no Rjo de Jan^o e os [deixei](#) uindome (- 75 - - A00_0065.
go, rezão hé que vos aqueiteis, mas não [deixeis](#) de proseguir adiante, pois
da Aldea, os quais ainda que alguns não [deixem](#) a vida viciosa por exempl
o que tem e sabe. 6. Gonçalo Alvarez: — [Deixemos](#), isto! Sou tãõ descuida
e não demenua ha criação do gado que lá [deixey](#), 5. E ha terra que á-de pe
oas de boca e ou tros m^tos furiozos q' [deixo](#) e grandes. Por estes rios acha
hum dos quais navegação caravelões, etc., [deixo](#) para o P.ºe João de Mello, r
isse, posto que com muyto trabalho. 18. [Deixo](#) de contar de outras ynfermid
e, e disto direi abaixo mais largo. 28. [Deixo](#) de dizer hum geral açoute, que
tamanha carta como esta e não acabaria. [Deixo](#) tambem porque lá vai a ca

C:\Documents and Settings\sandra.SANDRA1\Desktop\Trabalho_Unitex\Historical Portuguese (Brazil)\Corpus\corpus_milenio_snt\concord.html

0 sentence delimiter, 10805765 (186851 diff) tokens, 4906107 (170169) simple forms, 121578 (10) digits
4545174 occurrences (95551 DLF entries) simple words, 479 occurrences (424 DLC entries) compound words

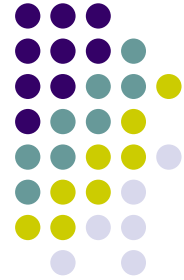
4. E o mesmo aviso se devia dar hà. Baya ao Padre Luis da Grãa para que acrecente e não demenua ha c
[deixey](#).

5. E ha terra que á-de pedir a Martim Afonso hé esta: scilicet, ao longo do mar do Rio de Yguape até o Rio
pouco mais ou menos de costa, e pera o sertão 3 ou 4 legoas; e se Martim Afonso for propicio podem pedi
de Iguape tres ou quatro legoas ao longo do mar, e outras tantas pera o sertão de largura. E se for caso qu
nos enchão esta dada ao diante donde não estiver dado.

{A00_0053.txt,.N}{67. - LISBOA AGOSTO-SETEMBRO DE 1561 391 - A00_0053.txt,.N}

67

Busca na lista de desconhecidas no Unitex



Unitex 2.0beta (July 19, 2007) - current language is Historical Portuguese (Brazil)

Text DELA FSGraph Lexicon-Grammar Edit File Edition Windows Info

Word Lists in C:\Documents and Settings\sandra.SANDRA1\Desktop\Trabalho_Unitex\Historical Portuguese (Brazil)\Corpus\co

DLF: 95551 simple-word lexical entries

deixa,deixar.V:P3s:Y2s
deixação,deixam.N+VAR:ms
deixação,deixam.N+VAR:ms
deixada,deixado.A:fs
deixada,deixar.V:K
deixadas,deixado.A:fp
deixadas,deixar.V:K
deixado,.A:ms
deixado,deixar.V:K
deixados,.N+VAR:ms
deixados,deixado.A:mp
deixados,deixar.V:K
deixaes,deixais.N+VAR:ms
deixai,deixar.V:Y2p
deixais,.N+VAR:ms
deixais,deixar.V:P2p
deixalla,deixala.N+VAR:ms
deixallas,deixalas.N+VAR:ms
deixallas,deixalas.N+VAR:ms

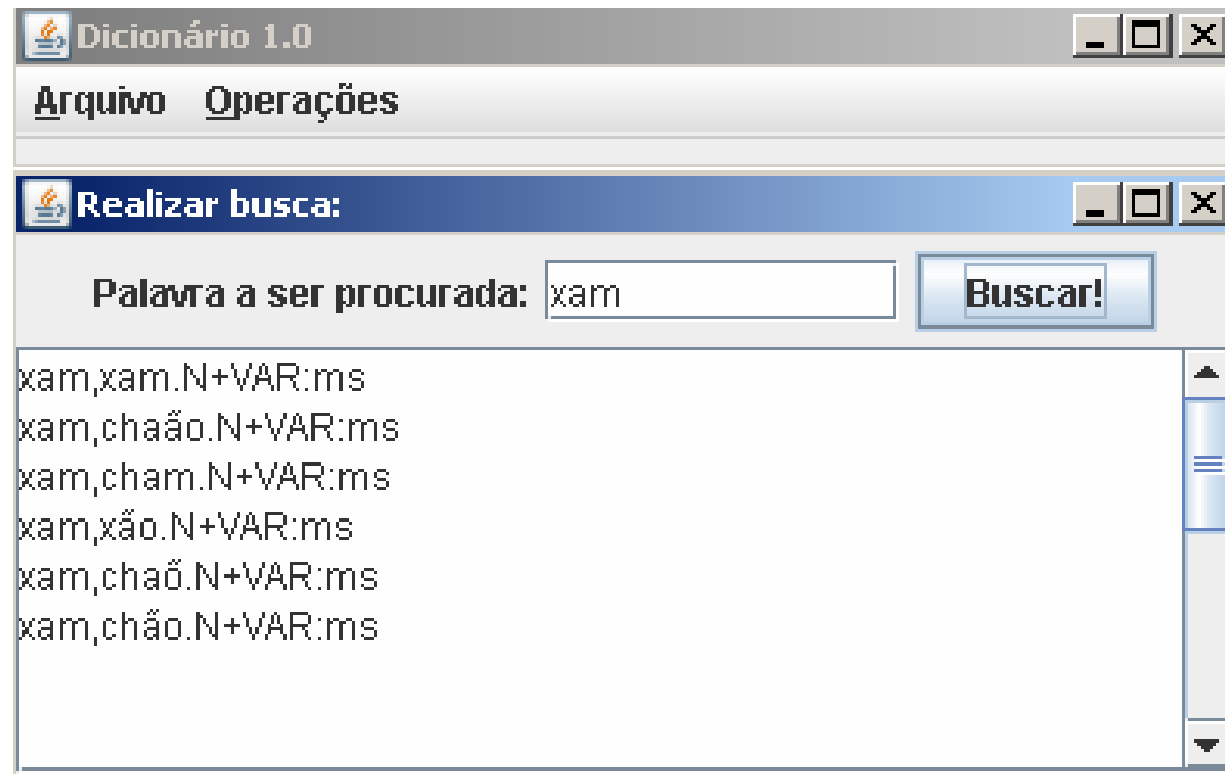
ERR: 80122 unknown simple

deixá
dèixa
deixássemos
deixaar
deixaçe
Deixae
deixae
deixair
deixal
deixalloy
deixamo
deixámo
deixamol
deixandonos
deixandoos
deixará
deixára
deixará
deixarám
deixaráó
deixaráó
deixaráó
deixaráon
deixaráon
deixaremna
deixariamos
deixarieis
deixamos
deixasão
deixasem
deixassemos
deixaua
deixauaó
deixavamos

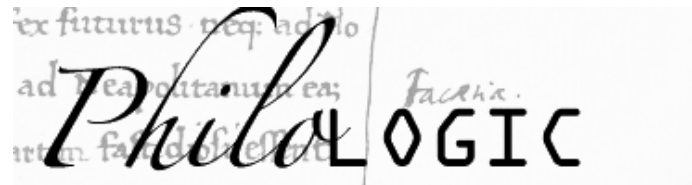
DLC: 424 compound lexical entries

abaixo-assinados,abaixo-assinado.N+ADVA:mp
ag\,agosto.N+ABREV:ms
ag\,agregada.A+ABREV:fs
ag\,agregada.N+ABREV:fs
ag\,agregada.V+ABREV:K
ag\,agregado.A+ABREV:ms
ag\,agregado.N+ABREV:ms
ag\,agregado.V+ABREV:K
ag\,agregados.A+ABREV:mp
ag\,agregados.N+ABREV:mp
ag\,agregados.V+ABREV:K
ag\,água.N+ABREV:fs
an\^tn anstn N+ARRFV:MS

Xam - SIACONF



Xam - Philologic



Welcome to PhiloLogic
[home](#) | [the ARTFL project](#) | [download](#) | [documentation](#) | [sample databases](#)

Found 24 matches, shown with frequencies in entire database.

Select words to search in the entire database. Select output options and bibliographic criteria below.

or

- 307 am
- 1 axam
- 151 cam
- 18 dam
- 41 eam
- 2 gam
- 32 ham
- 329 iam
- 50 jam
- 7 lam
- 41 mam
- 505 nam
- 15 pam
- 15 ram
- 533 sam
- 501 tam
- 7 uam
- 23 vam
- 5 xa
- 2 xam
- 6 xama
- 1 xas
- 1 yam
- 1 zam

Chão - Philologic



Welcome to PhiloLogic

[home](#) | [the ARTFL project](#) | [download](#) | [documentation](#) | [sample databases](#)

Found 13 matches, shown with frequencies in entire database.

Select words to search in the entire database. Select output options and bibliographic criteria below.

or

361	<input type="checkbox"/>	achão	
2	<input type="checkbox"/>	chaão	█
1	<input type="checkbox"/>	chião	
2	<input type="checkbox"/>	chã	
6	<input type="checkbox"/>	chãa	
284	<input type="checkbox"/>	chão	█
44	<input type="checkbox"/>	chãos	
1	<input type="checkbox"/>	chãs	
2	<input type="checkbox"/>	coão	
1	<input type="checkbox"/>	crão	
117	<input type="checkbox"/>	cão	
1	<input type="checkbox"/>	ehão	
979	<input type="checkbox"/>	hão	

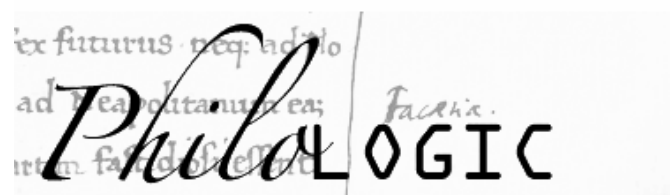
Lista de desconhecidas

Historical Portuguese (Brazil)\Corpus\corpus_milenio_snt

ERR: 80122 unknown simple words

chanaa
Chanaan
Chanão
chanas
chancel
chanchalaria
chanchonetas
chançonetas
Chandues
Chanés
chaneza
chans
chansellaria
chanseller
Chances
Chantre
chantre
chaó
chaos
cháos
Chaparé
Chapatel
chapèo
chapéios
Chapéo
chapeo
chapéo
Chapéos
chapéos
chapi
chapim
chapins
chapinz
chapidéo

Checagem no Philologic: cháó



Welcome to PhiloLogic

[home](#) | [the ARTFL project](#) | [download](#) | [documentation](#) | [sample databases](#)

Your search found **1** occurrences

[Click here for a KWIC Report](#)

Occurrences 1-1:

1. JOSEPH BARBOZA DE... / JOSEPH BARBOZA DE SÁ... [[Paragraph](#) | [Section](#)]

Villa com os augmentos que a possibilidade dos moradores permite mas sem torre por se lhe não fazer em sua erecção e oppondose depois a hisso o Vigario e Padre Manoel anno mil sette centos cincoenta e cinco a sua custa e do Povo estando ja a Torre em boa altura cahio no **cháó** por erro que de seu principio levou querendo neste anno em que \ intento o Doutor Joze Pereira Duarte o não fazia por falta de pessoa intelligente que mestrasse a obra porque lhe não acontecesse como ao seu Predecessor. E como todos os humanos

Processadores para *córpus históricos*



- Unitex é uma implementação livre do programa Intex, ambos criados no laboratório francês LADL (Laboratoire d'Automatique Documentaire et Linguistique).
 - Os dicionários *Unitex* se baseiam no formalismo DELA (*Dictionnaire Electronique du LADL*) também desenvolvido no laboratório LADL.
 - O suporte ao idioma português é particularmente bom graças ao trabalho Unitex-PB desenvolvido em um mestrado do NILC.
 - <http://www-igm.univ-mlv.fr/~unitex/> e <http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html>
- Philologic é uma ferramenta para buscas avançadas em corpus desenvolvida pelo projeto ARTFL (American and French Research on the Treasury of the French Language) na universidade de Chicago.
 - <http://humanities.uchicago.edu/orgs/ARTFL/>



*Processadores para **córpus históricos***

Recurso	Philologic	Unitex
Execução	Remota (Web)	Local (janelas)
Anotação	XML-TEI	Gramatical, sentencial
Subcórpus	Sim	Não
Buscas avançadas	Bibliografia, colocações	Léxicos

Pesquisa: criar um sistema com o melhor dos 2 mundos

O **Unitex**

fornece buscas poderosas e acesso a léxicos;
instalação ao alcance de todos
foi personalizado para trabalhar com o alfabeto do Português Histórico

O **Philologic**

tem como ponto forte a facilidade de uso (Web),
centralização e suporte a texto anotado em um padrão internacional;
instalação exige especialista em computação

Unitex usa UNICODE (UTF-16) e o **Philologic** UNICODE (UTF-8)

Obrigada!



Material do Curso sobre C3rpus Hist3ricos & DHPB:

<http://moodle.icmc.usp.br/ebral/>

Referências



Sanderson, Robert; "Historical Text Mining", Historical "Text Mining" and "Historical Text" Mining: Challenges and Opportunities. Talk presented at *Historical Text Mining Workshop*, July 2006, Lancaster University, UK. (Available at: <http://ucrel.lancs.ac.uk/events/htm06/>)

Rydberg-Cox, Jeffrey A. 2003. Automatic disambiguation of Latin abbreviations in early modern texts for humanities digital libraries. In: *Proceedings of JCDL*, 03, p. 372-373.