O papel do léxico do Unitex-PB no projeto "Dicionário Histórico do Português do Brasil dos séculos XVI, XVII e XVIII"

Sandra Maria Aluísio
(NILC-ICMC-USP)

Projeto - Institutos do Milênio (3 anos)

- **18 professores doutores e alunos de graduação e mestrado**

- **Instituição-sede:**
- FCL da UNESP, Araraquara (**coordenadora Profa. M. T. Biderman**)
  Instituições parceiras:
- Universidade de Évora, Portugal
  Universidade de São Paulo, Campus de São Paulo e de São Carlos
  Universidade Federal de São Carlos
  Universidade Federal do Rio Grande do Sul
  Universidade Federal de Minas Gerais
  Universidade Federal do Mato Grosso do Sul
  Universidade Federal da Bahia
  Universidade Federal de Uberlândia
  Universidade Federal do Rio de Janeiro
  Universidade Estadual de Londrina

Objetivo

- 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)

Objetivos Pontuais: o córpus

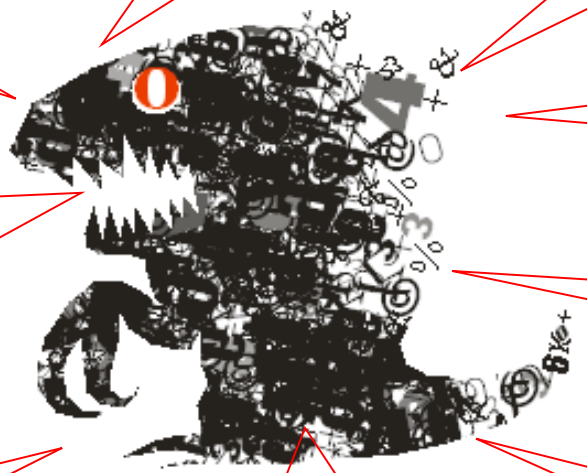
- Criar um córpus de referência bastante representativo dos séculos em questão para embasar adequadamente o dicionário.
  - Hipótese inicial: criar um corpus de 3.000.000 de palavras (três milhões)
  - Textos sobre o Brasil e produzidos por brasileiros, ou portugueses radicados definitivamente no país, para permitir a recuperação do repertório vocabular usado nos séculos XVI, XVII e XVIII.
- “Isso feito, e sistematizado em forma de dicionário, poderemos divulgar os resultados desta pesquisa para os brasileiros “leigos” poderem consultar.” (Biderman, projeto)

O córpus

- 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
  - As fontes de referência estarão em parte **publicadas**, e, no que concerne às fontes manuscritas, serão também perfeitamente identificadas.
  - Para podermos ter uma base textual informatizada de dimensões relativamente grande é preciso planejar a informatização para o período de **um ano**.
  - Posteriormente após ser produzido o protótipo de **10.000 verbetes** começaríamos a refinar a qualidade da informação.

(Biderman, projeto)

Contando uma história...



Tipologia textual

Tratamento da Hifenização

Tratamento de abreviaturas

Variação da grafia

Caracteres acentuados ou não que não pertencem ao latim básico ou estendido

Amostra representativa para a tarefa: léxico

Regras de Transcrição e Notas do editor

Metadados para o cabeçalho dos textos

Processador de córpus

Embora façamos uso de textos com intervenção dos editores ...

- Separação de palavras que no manuscrito estão grudadas [aestimavel cartade= a estimavel carta de; deque porculpa do patraõ, ePratico= de que por culpa do patraõ, e Pratico];

- Introdução de pontuação inexistente no manuscrito bem como paragrafação para ajudar no entendimento do texto.

“ 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)

Há ainda alguns monstros

.... de Britto

..... de Britto

Auto de enventar[io] que o juis ordi[nário e dos] or[fãos] ant° Correia da silva mãodou fazer por falesimento de fr^co bicudo de britto

1650 Nº 44

Muitas abreviaturas

.....

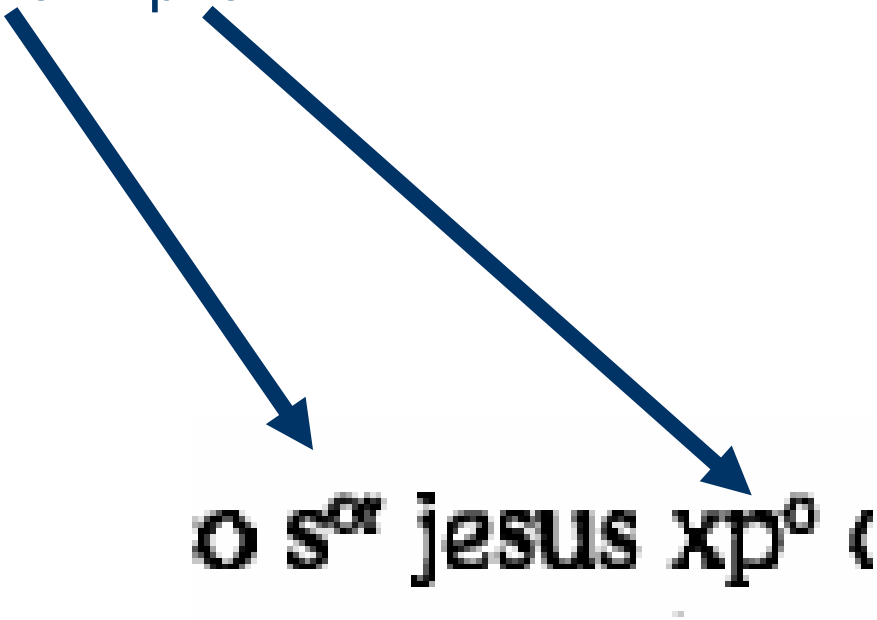
Anno de nasimento de nosso s^or jesus xp° de mil e seis sentos e sincoenta e quatro annos en os trinta dias do mes de marsso da sobredita era nesta vila de santa anna da parnaiba da cap^ta de são v^te estado do brazil Ett^a. nesta dita vila nas cazas da morada que foi de fr^co bicudo de brito que ds ten pelo juis ordinario e dos orfãos ant° correia da silva foi mãodado a min t^am e escrivão fazer este auto p^a por ele eventariar os bẽis e fazenda que ficou por morte e falesimẽto de fr^co bicudo de brito que dõs tẽ p^a o que deu juramento dos santos evangelhos a viuva tomazia Ribr^a mulher que foi do dito defunto p^a que sob cargo dele declarasse e manifestasse todos os bẽis e fazenda que pesuhia asin moveis como de rais dr° 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), SÍLNIA NUNES MARTINS, EDITORA RESPONSÁVEL DA DIVISÃO DE ARQUIVOS DO ESTADO DE SÃO PAULO

Anotação de adição, omissão, correção do Editor

Formas das Abreviaturas já pré-processadas

- sarg.^to P.^e S.^or S.^r m.^to grd.^e dr.^o
- q^m P^e I^o V^te s^or xp^o
- @
- 8.bro
- Carv. q. Sr.
- Sñor



Temos que processar essas anotações

O padre noviço, que acompanhou ao Padre Francisco Veloso, teve mais bom [tempo ?] de experiência nesta pere-grinação, porque além da fome, que a caridade fez volun-tária 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 |

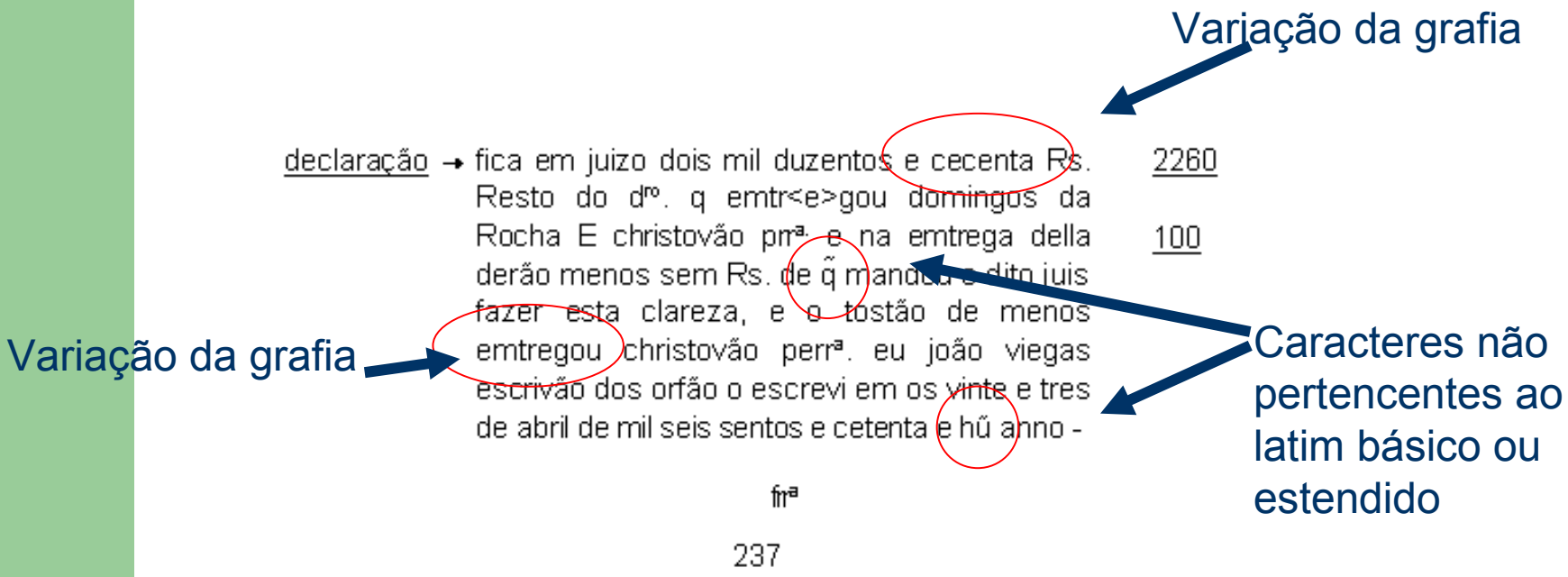
CARTA LXVI - AO PADRE PROVINCIAL DO BRASIL
1654, ANTÓNIO VIEIRA , J. LÚCIO D'AZEVEDO (ed.)

Anotação pelo Editor de partes não legíveis

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

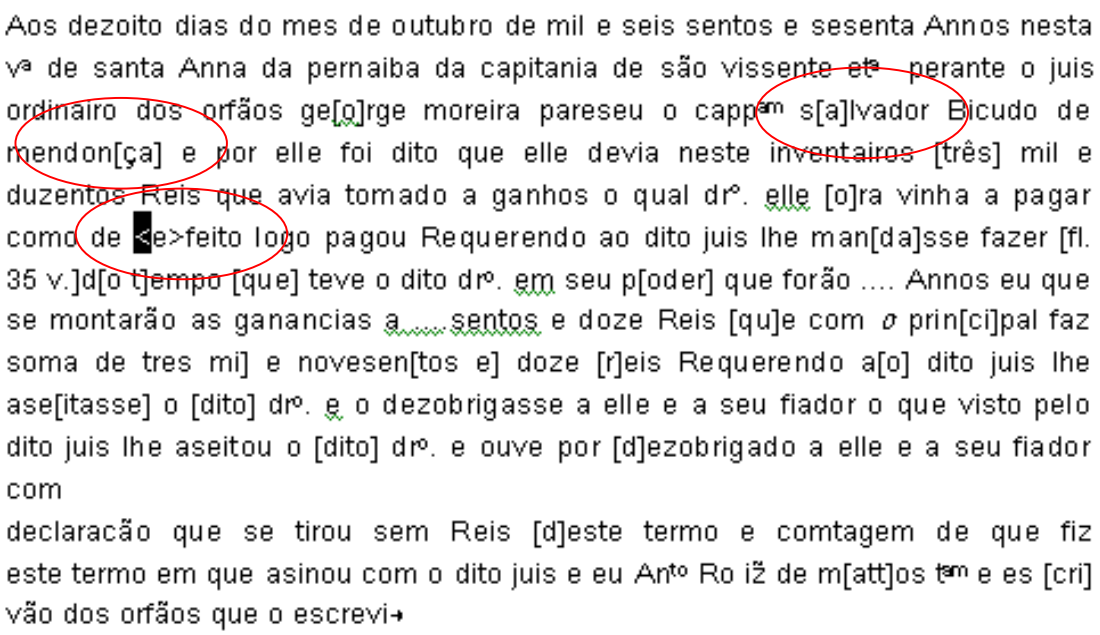
CARTA XVII - AO MARQUÊS DE NIZA 1648 — JANEIRO 12,
ANTÓNIO VIEIRA , J. LÚCIO D'AZEVEDO (ed.)

Temos que processar essas anotações



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

Temos que processar essas anotações



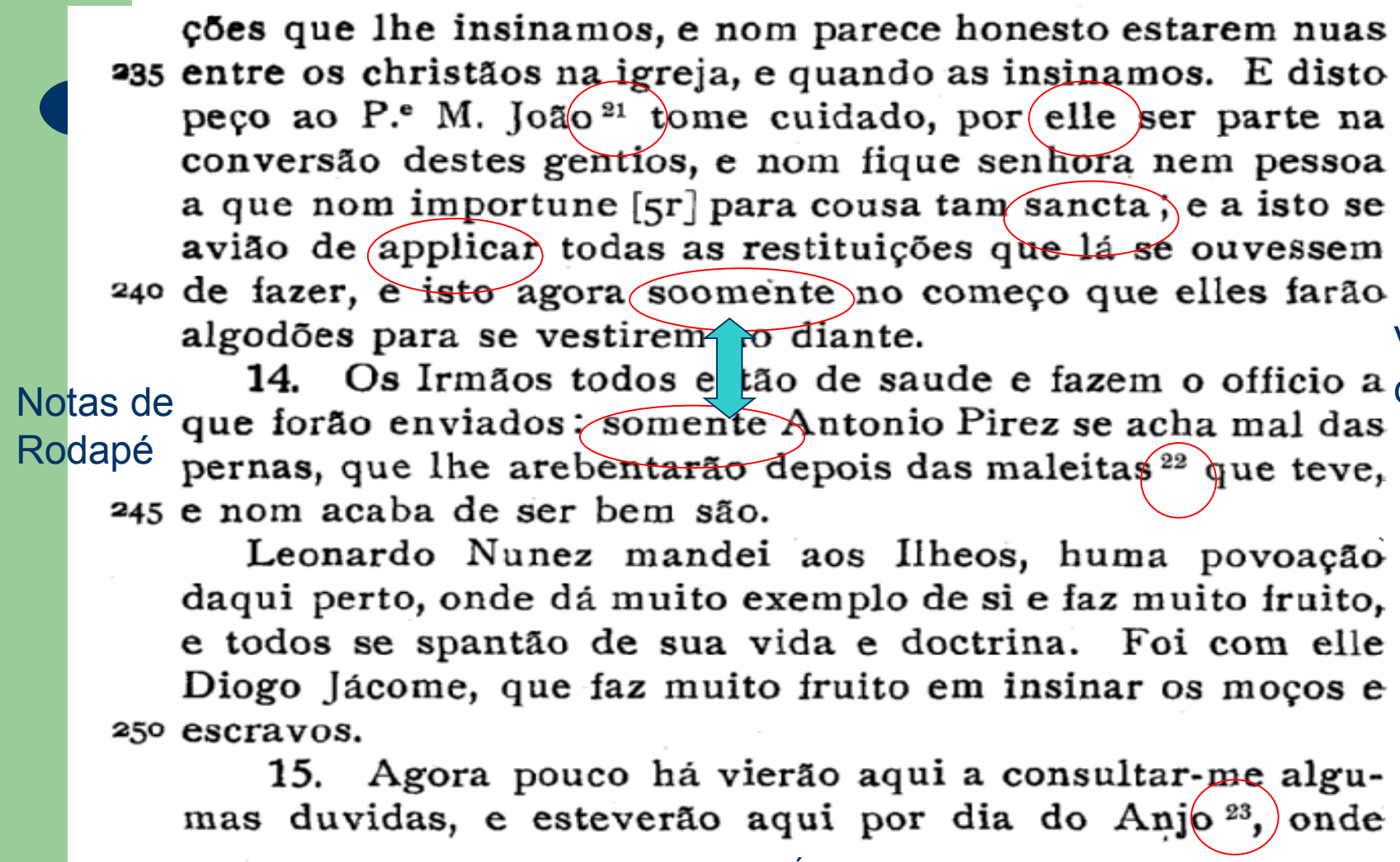
Aos dezoito dias do mes de outubro de mil e seis sentos e sesenta Annos nesta vª de santa Anna da pernaiba da capitania de são vissente etª perante o juis ordinairo dos orfãos ge[o]rge moreira pareseu o cappam s[a]lvador Bicudo de mendon[ça] e por elle foi dito que elle devia neste inventairos [três] mil e duzentos Reis que avia tomado a ganhos o qual drº. elle [o]ra vinha a pagar como de <e>feito logo pagou Requerendo ao dito juis lhe man[da]sse fazer [fl. 35 v.]d[o t]empo [que] teve o dito drº. 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] drº. e o dezobrigasse a elle e a seu fiador o que visto pelo dito juis lhe aseitou o [dito] drº. e ouve por [d]ezobrigado a elle e a seu fiador com

declaracão que se tirou sem Reis [d]este termo e comtagem de que fiz este termo em que asinou com o dito juis e eu Antº Ro iž de m[att]os tam e es [cri]vão dos orfãos que o escrevi→

Mistura de padrões de anotação do Editor

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

Temos que processar essas anotações



ções que lhe insinamos, e nom parece honesto estarem nuas
entre os christãos na igreja, e quando as insinamos. E disto
peço ao P.e M. João [21] 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 soomente no começo que elles farão
algodões para se vestirem [illegible]o 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 [22] 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.

15. Agora pouco há vierão aqui a consultar-me algu-
mas duvidas, e esteverão aqui por dia do Anjo [23], onde

Notas de Rodapé

Variações de grafias

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

Mais variações de grafia complicando a contagem da freqüência de palavras do córpus ...

que lhe **insinamos**, e **nom** parece honesto estarem nuas entre os **christãos** na igreja, e quando as **insinamos**. E disto peço ao P.^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 **soomente** 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.)

Estágios da compilação de um córpus

- **projeto do córpus**, que inclui a seleção dos textos e os cuidados com os requisitos como
  - autenticidade, representatividade, balanceamento, amostragem, diversidade e tamanho
- **compilação** (ou captura), manipulação, nomeação dos arquivos de textos, e pedidos de permissão de uso, e
- **Anotação** tanto estrutural como lingüística.

Relatório Parcial do Projeto - Institutos do Milênio - proc. 420139/2005-2

"A **identificação e a localização das obras** e dos documentos que constituirão as fontes de referência do DICIONÁRIO HISTÓRICO DO PORTUGUÊS DO BRASIL (sécs. XVI,XVII e XVIII) **constitui parte central de nosso projeto**, seu núcleo essencial, sendo seu ponto de partida. Dependendo da qualidade, variedade e representatividade dos textos que conseguirmos coletar e informatizar, tal será a qualidade do produto que vamos criar, isto é, o *dicionário*."

...

"Por outro lado, concluímos 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."

Anotação

- *Anotação estrutural:* marcação de **dados externos** e **internos** dos textos.

- 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..

Projeto do Córpus

Metadados para o cabeçalho dos textos

Tipologia textual

Variação da grafia

Amostra representativa para a tarefa: léxico histórico

Compilação e Anotação

Tratamento de abreviaturas

Codificação de caracteres

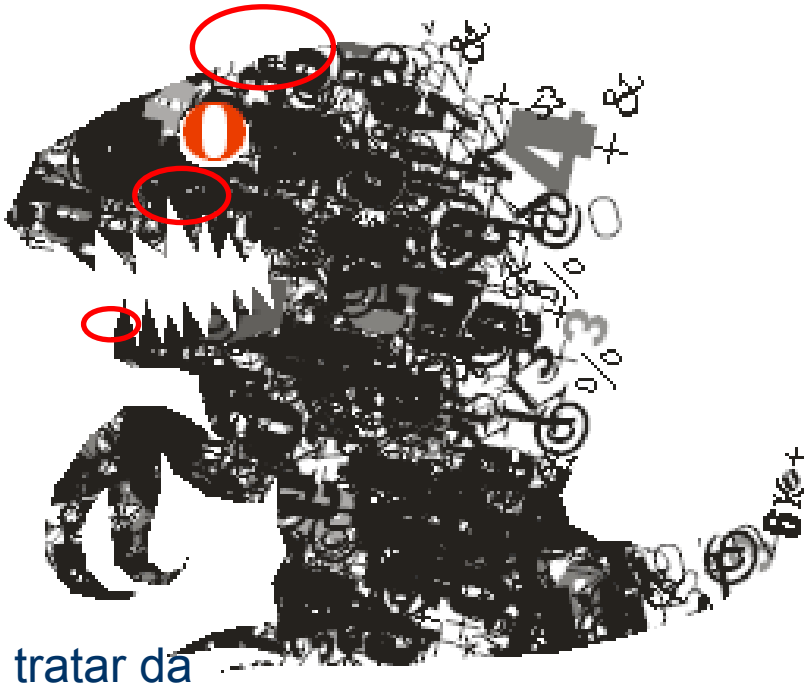
Processador de córpus

Transcrição, Notas do editor & tratamento da hifenização

Coordenar o trabalho de uma grande equipe

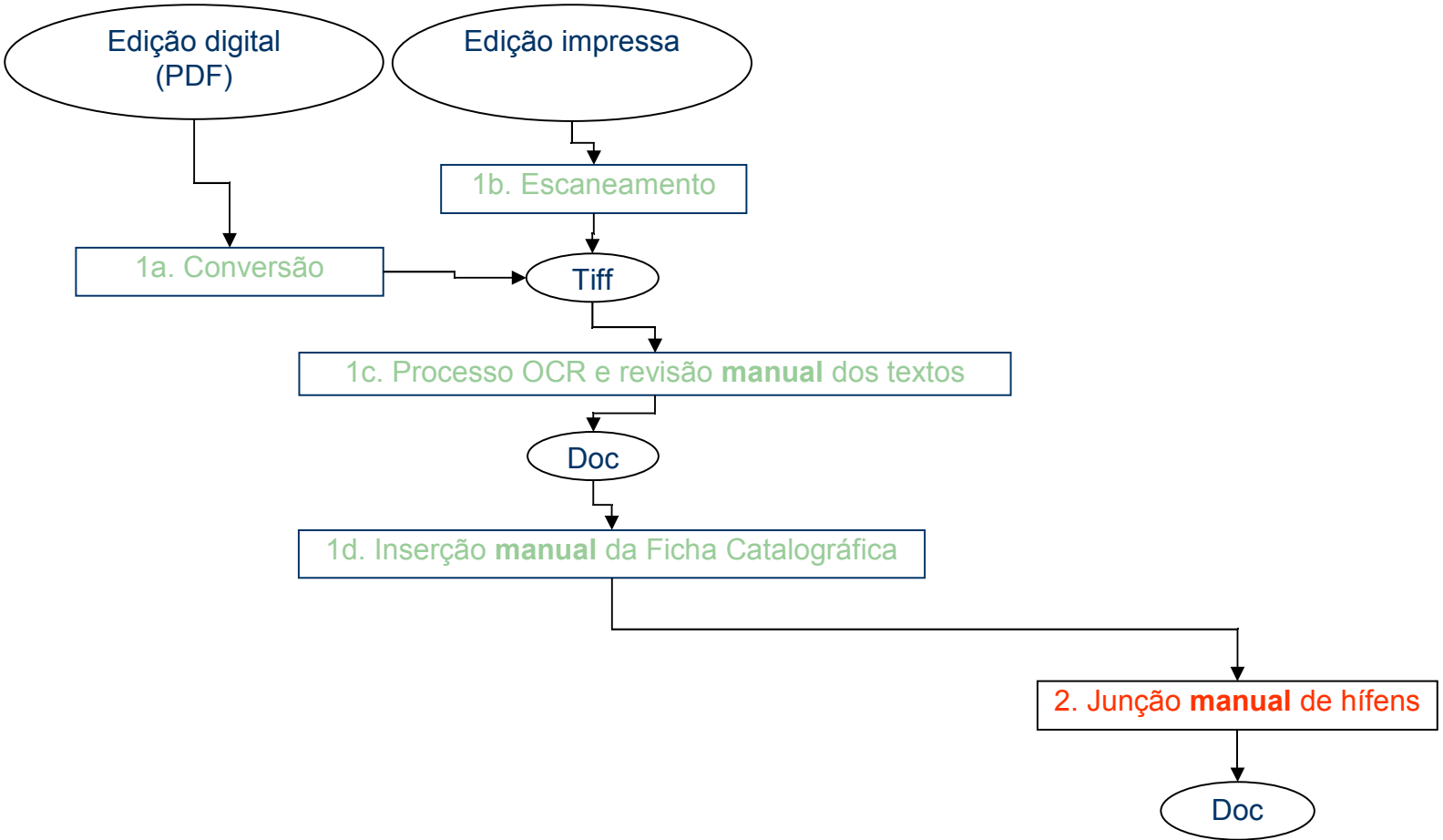
....criar um córpus de textos históricos é uma empreitada cara e demorada, portanto este tipo de córpus deve ser reusado por outros grupos de pesquisa e/ou outros projetos

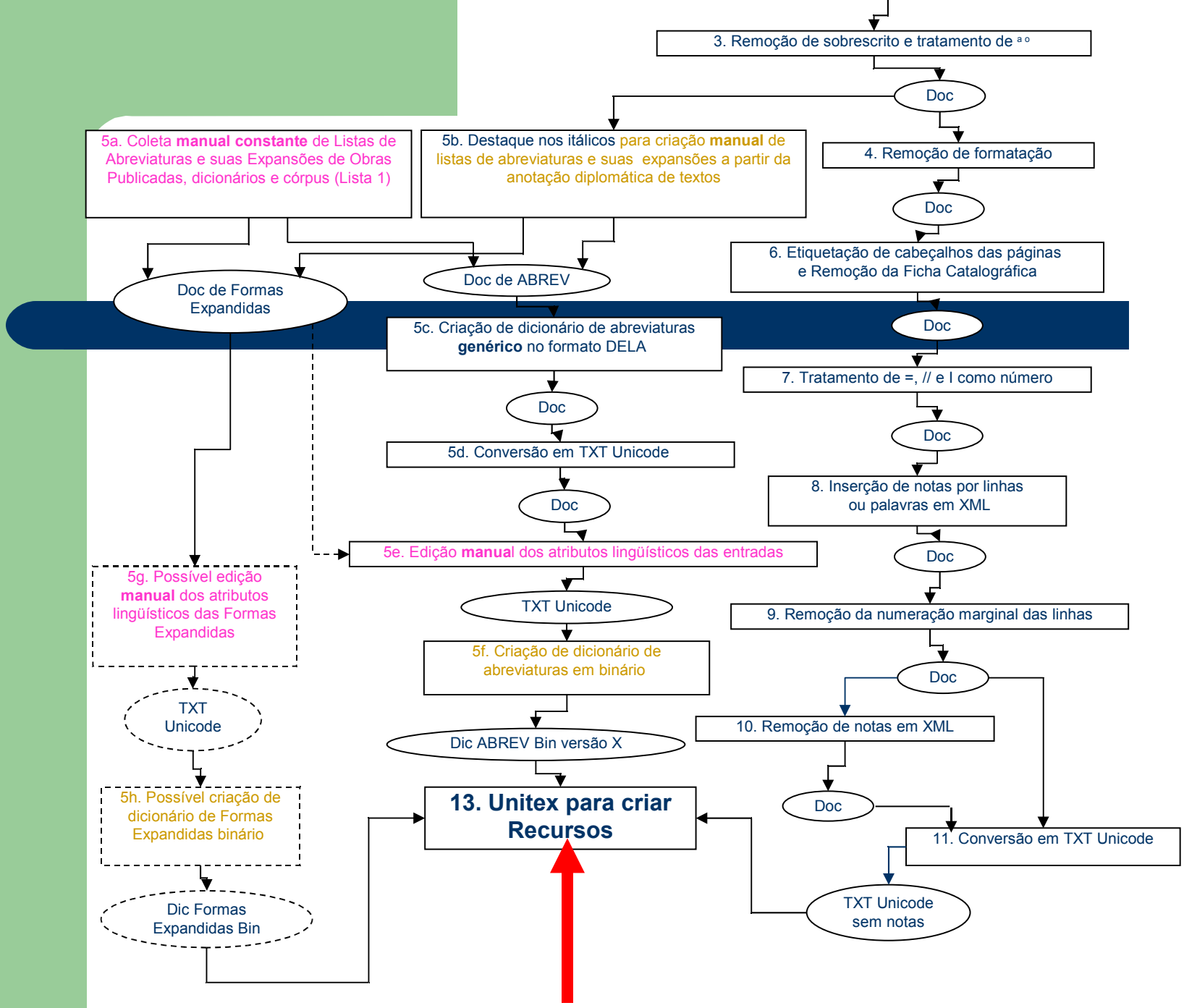
- Faz a seleção dos textos que comporão o córpus
- Escaneia e corrige erros de OCR
- Preenche cabeçalho com vários
- metadados
- Trata hifenização
- Pré-processa os textos para serem usados por processadores de córpus



- Adapta processadores de córpus para tratar da escalabilidade e funcionalidades adequadas à tarefa
- Anota fenômenos lingüísticos com padrões internacionais para que o córpus possa ser útil para outros projetos

O processo de codificação dos textos do Córpus: para ser utilizado com Unitex, Philologic e disponibilizado para outras pesquisas




3. Remoção de sobrescrito e tratamento de ª º
Doc
5a. Coleta manual constante de Listas de Abreviaturas e suas Expansões de Obras Publicadas, dicionários e córpus (Lista 1)
5b. Destaque nos itálicos para criação manual de listas de abreviaturas e suas expansões a partir da anotação diplomática de textos
4. Remoção de formatação
Doc
6. Etiquetação de cabeçalhos das páginas e Remoção da Ficha Catalográfica
Doc de Formas Expandidas
Doc de ABREV
Doc
5c. Criação de dicionário de abreviaturas genérico no formato DELA
7. Tratamento de =, // e I como número
Doc
Doc
5d. Conversão em TXT Unicode
8. Inserção de notas por linhas ou palavras em XML
Doc
Doc
5e. Edição manual dos atributos lingüísticos das entradas
5g. Possível edição manual dos atributos lingüísticos das Formas Expandidas
TXT Unicode
9. Remoção da numeração marginal das linhas
5f. Criação de dicionário de abreviaturas em binário
Doc
TXT Unicode
10. Remoção de notas em XML
Dic ABREV Bin versão X
5h. Possível criação de dicionário de Formas Expandidas binário
13. Unitex para criar Recursos
Doc
11. Conversão em TXT Unicode
Dic Formas Expandidas Bin
TXT Unicode sem notas

Roteiro

- O projeto - *Dicionário Histórico do Português do Brasil dos séculos XVI, XVII e XVIII*

- Desafios
  - Construir o Córpus
  - Construir o Córpus
  - Construir o Córpus
  - Construir o Córpus
  - Escolher & Adaptar um Processador de Córpus

- Unitex-PB e o Dicionário de **Abreviaturas**

- Metodologia para Detecção e Agrupamento de palavras com **grafias distintas**
  - para fornecer uma contagem próxima da real da freqüência de palavras do córpus – o papel da normatização
  - para informar a variação de grafia (um dos campos do dicionário e ajudar a eleger a entrada)

Escolha do Unitex-PB (Arnaldo C. mestrado)

- *Xaira* significa *XML Aware Indexing and Retrieval Architecture.*
  - A ferramenta foi construída como uma versão melhorada do software SARA – (*SGML Aware Retrieval Application*) criado pelo grupo do BNC.
  - http://www.oucs.ox.ac.uk/rts/xaira/

- *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 (*Dictionnarie 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/

Escolha do Unitex-PB

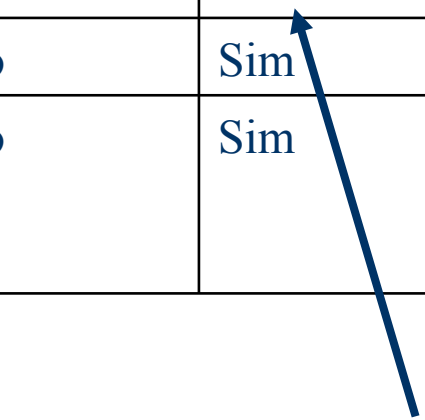
Recursos oferecidos pelas ferramentas

Aplicativo	Interface	Texto anotado	Portabilidade	Dicionário	Indexação	Subcorpus	Codificação de caracteres
Unitex	Janelas (Java)	Sim, somente etiquetas léxicas, num formato particular ao Unitex	Ambientes Java	**Sim**	Sim	Não	Unicode (UTF-16)
Xaira	Janelas (GTK)	TEI, XCES e formatos baseados em XML	Ambientes Windows e Unix	Não	Sim	Sim	Unicode (UTF-16)
Philologic	Web	TEI-Lite	Ambientes Web	Não	Sim	Sim	Unicode (UTF-8)

Escolha do Unitex-PB

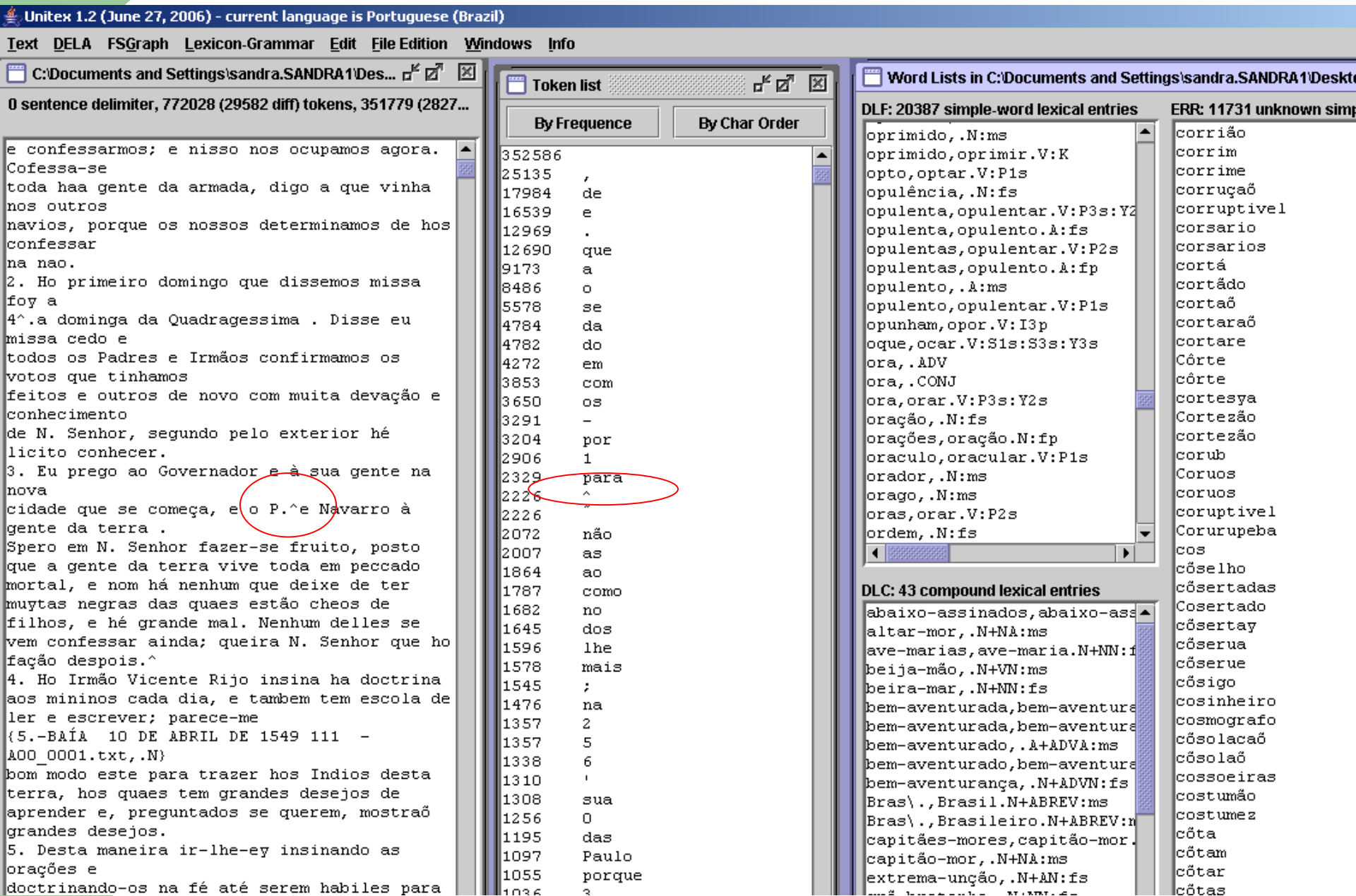
Buscas oferecidas pelas ferramentas

Aplicativo	Palavras simples	Frases	Expressões regulares	Classes gramaticais	Metadados
Unitex	Sim	Sim	Sim	Sim	Não
Xaira	Sim	Sim	Sim	Não	Sim
Philologic	Sim	Sim	Sim	Não	Sim



Adaptaremos o Unitex

Unitex-PB


Unitex 1.2 (June 27, 2006) - current language is Portuguese (Brazil)
Text DELA FSGraph Lexicon-Grammar Edit File Edition Windows Info
C:\Documents and Settings\sandra.SANDRA1\Des...
0 sentence delimiter, 772028 (29582 diff) tokens, 351779 (2827...
e confessarmos; e nisso nos ocupamos agora. Cofessa-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 dominga 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
de N. Senhor, segundo pelo exterior hé licito conhecer.
3. Eu prego ao Governador e à sua gente na nova
cidade que se começa, e o P.^e Navarro à gente da terra .
Spero em N. Senhor fazer-se fruito, posto que a gente da terra vive toda em peccado mortal, e nom há nenhum que deixe de ter muytas negras das quaes estão cheos de filhos, e hé grande mal. Nenhum delles se vem confessar ainda; queira N. Senhor que ho fação despois.^
4. Ho Irmão Vicente Rijo insina ha doctrina aos mininos cada dia, e tambem tem escola de ler e escrever; parece-me
{5.-BAÍA 10 DE ABRIL DE 1549 111 - A00_0001.txt,.N}
bom modo este para trazer hos Indios desta terra, hos quaes tem grandes desejos de aprender e, preguntados se querem, mostraõ grandes desejos.
5. Desta maneira ir-lhe-ey insinando as orações e
doctrinando-os na fé até serem habiles para
Token list
By Frequence
By Char Order
352586
25135 ,
17984 de
16539 e
12969 .
12690 que
9173 a
8486 o
5578 se
4784 da
4782 do
4272 em
3853 com
3650 os
3291 -
3204 por
2906 1
2329 para
2226 ^
2226 "
2072 não
2007 as
1864 ao
1787 como
1682 no
1645 dos
1596 lhe
1578 mais
1545 ;
1476 na
1357 2
1357 5
1338 6
1310 '
1308 sua
1256 0
1195 das
1097 Paulo
1055 porque
Word Lists in C:\Documents and Settings\sandra.SANDRA1\Deskt
DLF: 20387 simple-word lexical entries
oprimido,.N:ms
oprimido,oprimir.V:K
opto,optar.V:P1s
opulência,.N:fs
opulenta,opulentar.V:P3s:Y2
opulenta,opulento.A:fs
opulentas,opulentar.V:P2s
opulentas,opulento.A:fp
opulento,.A:ms
opulento,opulentar.V:P1s
opunham,opor.V:I3p
oque,ocar.V:S1s:S3s:Y3s
ora,.ADV
ora,.CONJ
ora,orar.V:P3s:Y2s
oração,.N:fs
orações,oração.N:fp
oraculo,oracular.V:P1s
orador,.N:ms
orago,.N:ms
oras,orar.V:P2s
ordem,.N:fs
DLC: 43 compound lexical entries
abaixo-assinados,abaixo-ass
altar-mor,.N+NA:ms
ave-marias,ave-maria.N+NN:f
beija-mão,.N+VN:ms
beira-mar,.N+NN:fs
bem-aventurada,bem-aventura
bem-aventurada,bem-aventura
bem-aventurado,.A+ADVA:ms
bem-aventurado,bem-aventura
bem-aventurança,.N+ADVN:fs
Bras\.,Brasil.N+ABREV:ms
Bras\.,Brasileiro.N+ABREV:n
capitães-mores,capitão-mor.
capitão-mor,.N+NA:ms
extrema-unção,.N+AN:fs
ERR: 11731 unknown simp
corrião
corrim
corrime
corruçaõ
corruptivel
corsario
corsarios
cortá
cortãdo
cortaõ
cortaraõ
cortare
Côrte
côrte
cortesya
Cortezão
cortezão
corub
Coruos
coruos
coruptivel
Corurupeba
cos
cõselho
cõsertadas
Cosertado
cõsertay
cõserua
cõserue
cõsigo
cosinheiro
cosmografo
cõsolacaõ
cõsolaõ
cossoeiras
costumão
costumez
cõta
cõtam
cõtar
cõtas

Dicionário de Abreviaturas

B,bastarda.N+ABREV:fs/sec18
B,bastarda.A+ABREV:fs/sec18

Abreviatura,expansão.ClasseGramatical+ABREV:atributos/comentários

- Tratamos a ambigüidade categorial
- **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

A^a,Aranha.N+ABREV:fs/sec19
A^a,Aranha.Npr+ABREV:ms/sec19
...
A^al,auxiliar.N+ABREV:ms/sec18
A^al,auxiliar.N+ABREV:fs/sec18
A^al,auxiliar.A+ABREV:ms/sec18
A^al,auxiliar.A+ABREV:fs/sec18
A^al,auxiliar.V+ABREV:W1s/sec18
A^al,auxiliar.V+ABREV:W3s/sec18
A^al,auxiliar.V+ABREV:U1s/sec18
A^al,auxiliar.V+ABREV:U3s/sec18

Clarissa G.Bengtson, Letras, UFSCar (Oto Vale) (IC)

Metodologia para Detecção e Agrupamento de palavras com grafias distintas

- **A abordagem tomada consiste em**
  - **aplicar uma série de regras de transformação ao córpus com o objetivo de agrupar grafias diferentes em torno de uma grafia comum.**
  - **Uma regra *transforma* uma grafia G1 se satisfaz às condições de cobertura da regra e produz a grafia G2, sempre de acordo com a sintaxe das regras de transformação.**
  - **Duas grafias G1 e G2 são *agrupadas* em torno de uma grafia G3 se uma coleção de regras produz a grafia G3 tanto para a grafia G1 quanto para a grafia G2.**

- **Baseada 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 normatização de textos históricos". *Dissertações do Instituto de Matemática e Estatística.* Universidade de São Paulo (2005)

Exemplos de Regras

Regras para grafias em desuso

y y i
ph ph f
ò ò ó
th th t

Regras para consoantes dobradas

[Menegatti, 02] observa a ocorrência de consoantes oclusivas e fricativas latinas dobradas. Tais consoantes dobradas podem ser removidas e substituídas por uma única ocorrência da mesma consoante.

ff ff f
pp pp p
tt tt t
cc cc c
bb bb b
dd dd d
gg gg g
vv vv v
zz zz z

Extensão da regra acima: Com base nessas sugestões e na observação de consoantes dobradas no córpus, as seguintes regras foram criadas:

mm mm m
nn nn n
ll ll l

Exemplos de Regras

Regras geradas de acordo com normas ortográficas

Regra	Aplica-se
m[cd...z] m n	“m” somente antes de “p” e “b”
n[pb] n m	“m” somente antes de “p” e “b”
aã aã ã	nasalisação ultrapassada
aõ aõ ão	nasaliação ultrapassada
aes$ e i	formas plurarizadas abandonadas (algodoaes: algodoais)

Exemplos de Regras

Regras baseadas em freqüência

Algumas regras foram criadas com o objetivo exclusivo de agrupar grafias, sem interesse em agrupá-las em torno de uma grafia moderna. Outras regras foram derivadas do trabalho de [Menegatti, 02]

Regra	**Agrupa**
chr chr cr	christa e crista; cristã e christã
ch ch x	cha e xa; debaixo e debaicho
ee ee é	maree e maré; neela e néla
pt pt t	promtamente, promptamente e prontamente
v$ v u	rev e reu
uu uu u	nuus e nus
j[bcd..xz] j i	acima e acjma; ainda e ajmda
.à à á	aliàs e aliás; cà e cá
ct ct t	exacto e exato; extincto e extinto
issim.?s?$ is is	digníssimo e digníssimo
mn mn n	emtregarão e emntregarão
mpt mp n	prompta e pronta
oens$ oen õe	proposiçõens e proposições
ozo$ z s	religiozo e religioso; rigorozo e rigoroso
[^r][aã]o$ [aã]o am	são, saão, são e sam; tão, taõ, taão e tam
e[oa] e ei	alldea, aldeia e aldea
^he he e	helle, hele, elle e ele

Exemplos de Regras

Regras lexicalizadas

Regra	Agrupa
deos o u	Deos e Deus; judeos e judeus

Regras automáticas

Em [Hirohashi, 05], regras de transformação são geradas de forma automática sobre o córpus Tycho Brahe. Algumas dessas regras são reutilizadas neste trabalho ao se mostrarem bastante eficientes para a tarefa de agrupação.

agio$ a á	suffragio, sufrágio e suffrágio
preciz z s	preciza e precisa
serviss ss ç	servisso e serviço
zente z s	prezente e presente
.acem$ c ss	tirassem e tiracem

Seqüência de regras aplicadas a uma mesma palavra

* PALAVRA CHAÕ

ch ch x	transforma "chaõ" em "xaõ"
[^aeiou]aõ aõ ão	transforma "xaõ" em "xão"
[^r][aã]o$ [aã]o am	transforma "xão" em "xam"

* PALAVRA CHAÃO

ch ch x	transforma "chaão" em "xaão"
aã aã ã	transforma "xaão" em "xão"
[^r][aã]o$ [aã]o am	transforma "xão" em "xam"

===> agrupamento de CHAÕ e CHAÃO em torno da grafia XAM

Seqüência de regras aplicadas a uma mesma palavra

* PALAVRA ACCOMMETTIDO
    tt tt t transforma "accommettido" em "accommetido"
    mm mm m transforma "accommetido" em "accometido"
    cc cc c transforma "accometido" em "acometido"

* PALAVRA ACOMMETTIDO
    tt tt t transforma "acommettido" em "acommetido"
    mm mm m transforma "acommetido" em "acometido"

* PALVRA ACCOMETTIDO
    tt tt t transforma "accomettido" em "accometido"
    cc cc c transforma "accometido" em "acometido"

* PALAVRA ACOMMETIDO
    mm mm m transforma "acommetido" em "acometido"

===> agrupamento de ACCOMMETTIDO, ACOMMETTIDO, ACCOMETTIDO e ACOMMETIDO em torno da grafia ACOMETIDO

Agrupamentos

xam:

xão	(1)
chão	(183)
chaõ	(2)
cham	(1)
chaão	(2)

setembro:

setenbro	(4)
septembro	(9)
settembro	(1)
septenbro	(4)
setembro	(238)

xamam:

chamam	(656)
chamão	(485)
chamaõ	(7)
chamao	(3)

Agrupamentos

vinham:

vynham	(1)
vinhao	(1)
vinhaõ	(2)
vynhão	(2)
vinhão	(92)
vinham	(146)

vila:

vyla	(12)
villa	(1652)
vila	(843)

trinta:

ttrintta	(1)
trimta	(4)
trymta	(2)
trinta	(524)

Agrupamentos

abaixo:

abayxo	(1)
abaicho	(2)
abaixo	(420)

aceitar:

aceytar	(1)
acceitar	(2)
aceitar	(47)

Agrupamentos

ainda:

ajnda	(10)
aynda	(30)
ajmda	(2)
aimda	(5)
ainda	(2482)

aipim:

aypim	(2)
aipim	(2)

algodões:

algodoens	(1)
algodões	(16)

Agrupamentos

aldeia:
- alldea (1)
- aldeya (2)
- aldea (142)
- aldeia (460)

aldeiados:
- aldeados (13)
- aldeiados (2)

aldeianos:
- aldeanos (8)
- aldeianos (4)

aldeiar:
- aldear (1)
- aldeiar (4)

aldeiarem:
- aldearem (1)
- aldeiarem (1)

aldeias:
- aldeyas (5)
- aldeas (109)
- aldeias (496)

Agrupamentos

asim:

asym	(2)
asim	(315)

assi:

assy	(97)
assi	(291)

até:

athé	(13)
atee	(14)
até	(3340)

cristandade:

christandade	(59)
cristandade	(27)

Agrupamentos

ditar:
- dictar (3)
- ditar (1)

ditas:
- dittas (29)
- dictas (14)
- ditas (588)

ditava:
- dictava (6)
- ditava (2)

dito:
- dicto (68)
- dyto (8)
- ditto (467)
- dito (4156)

ditos:
- dictos (19)
- dittos (29)
- dytos (1)
- ditos (712)

Sandbox

Beñs	(23)
Brazil	(526)
acções	(63)
admiravel	(36)
agazalho	(14)
agoas	(55)
agua	(598)
aguas	(362)
aseitado	(4)
aseitar	(8)
aseitasse	(2)
aseitava	(2)
aseitei	(1)
aseito	(8)
aseitou	(33)
ésta	(6)
éstas	(1)
êste	(808)
éste	(3)
êstes	(308)
vierao	(1)
vierão	(197)
vigario	(277)

Sandbox

aderemço	(1)
aderencias	(1)
adespejalo	(1)
adespeza	(1)
adevertencia	(8)
adevertice	(2)
adevertio	(4)
adevertir	(2)
adevinhador	(1)
adevinhava	(4)
adevinho	(4)
adevirta	(2)
adevirto	(16)
adezreis	(1)
adherencia	(2)
adherente	(1)
adiantá	(1)
adiantára	(1)
adicois	(6)
adientadas	(1)
adiente	(2)
adiministrada	(1)
adimirados	(6)
adimiravelm	(2)

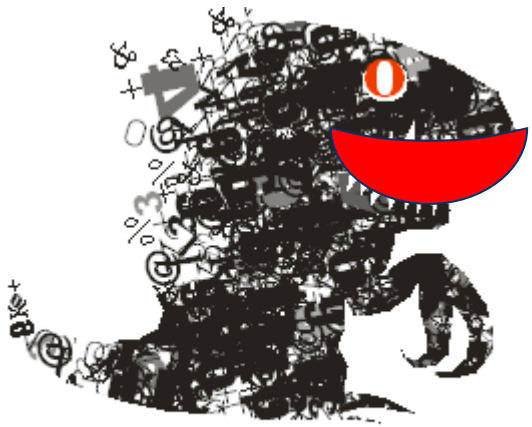
Sandbox

moedaz	(4)
moenga	(1)
moente	(6)
moentes	(3)
moeráõ	(1)
moẽ	(1)
mofama	(1)
mogadouro	(4)
mogé	(1)
mogî	(1)
mogí	(2)
moi	(1)
moida	(8)
moidas	(1)
moido	(3)
moidos	(2)
moieira	(1)
moio	(6)
moios	(17)
moitões	(1)

Sandbox

ocasiõis	(3)
ocasiois	(2)
ocasion	(2)
ocasionen	(2)
ocaziões	(2)
ocazioins	(2)
ocaziois	(8)
ocaziõis	(1)
ocaziona	(1)
ocazionis	(1)
ocazions	(6)
ocio	(7)
ocios	(1)
pargos	(3)
parianaz	(1)
paribus	(1)
paricatiba	(1)
paridura	(1)
parijó	(13)
parima	(17)
parimé	(6)
parime	(31)
paripatetica	(1)

Obrigada!



http://www.nilc.icmc.usp.br/