

*Acesso ao Córpus do Dicionário do
Português Histórico do Brasil:
As ferramentas*

Arnaldo Candido Junior

**Orientadora:
Sandra Maria Aluísio**

**Apoio:
Conselho Nacional de Desenvolvimento Científico e Tecnológico**

Roteiro

- Uso do **córpus** com **Philologic**
- Uso do **córpus** com **Unitex**
 - **Acesso a léxicos**
- **Buscas avançadas**
- **Proposta de um ambiente para a edição on-line de verbetes**

Estado atual do corpus

- 8 corpus digitalizados e processados
 - 4.160.807 palavras (MS Word)
 - 1203 textos
 - 46.4 Mb (codificação UTF-16)
- 9o. corpus sendo processado
- Ferramentas **Unitex** e **Philologic** prontas para extração de concordâncias

Estado atual do corpus (2)

- Pequenos problemas na extração automática de dados catalográficos
 - Data de edição usada como data de criação em alguns textos
 - “marcadordepagina” e “marcadorrodape” no texto

Estado atual do corpus (3)

- Pequenos erros de digitalização
 - Troca entre zero e “o” maiúsculo; entre um, “i” maiúsculo e “L” minúsculo
 - Caracteres inválidos (copyright por exemplo)

Acesso ao Philologic

- <http://moodle.icmc.usp.br/milenio>

Busca no projeto do Dicionário Histórico do Português do Brasil - Iceweasel

File Edit View History Bookmarks Tools Help

http://moodle.icmc.usp.br/philologic/

Google

Projeto do Dicionário Histórico do Português do Brasil

Buscar concordâncias ou documentos

Procurar por:

frecha [Buscar] [Limpar]

Mostar: Contexto KWIC Busca por similaridade

Contexto da Busca:

Palavra ou frase Frase separada por 3 palavras

Tipo de busca por proximidade: Sentença Parágrafo

Busca por dados bibliográficos:

Título: [] (exemplo: 'DO CLIMA E TERRA DO BRASIL')

Autor: [] (exemplo: 'TOMÉ DE SOUSA')

Date: [] (exemplo: '1551')

[Mais dados bibliográficos] [Resultados de busca refinados] [Busca por objetos de texto (capa, cabeçalho)] [Informações e ajuda (em inglês)]

POWERED BY PhiloLOGIC RTFL/DLDC/UCHICAGO

Find: word padding [Next] [Previous] [Highlight all] [Match case] [Reached end of page, continued from top]

Done

Busca no corpus

Philologic Results - Iceweasel

File Edit View History Bookmarks Tools Help

http://moodle.icmc.usp.br/cgi-bin/newphilo/search3t?dbname=miler Google

This page contains the first 25 occurrences. Please follow the link(s) at the bottom of the page to see the rest of the occurrences your search found.

1. P. MANUEL DA... . *CARTA DO P. MANUEL DA NÓBREGA A...* [Paragraph | Section]
e mattarão dous ou tres homens que acharão no caminho dos Ylheos pera Porto Seguro e derão em huma roça de christãos nos Ylheos. E passando pollo Engenho de S. Joam, em que estava Thomás Alegre, meteo N. Senhor tanto medo nos ossos dos christãos, que despovoão o Engenho sem yndio atirar **frecha**, antes se cree que, já satisfeitos da morte dos seus, se contentavão, porque a muytos christãos que poderão matar e roubar muy liberalmente deixarão yr. Como isto se soou, entrou o mesmo medo nos outros Engenhos e sem verem yndio despovoão e alargão tudo, recolhendo-se na Vila; o
2. P. LEONARDO DO VALE. *CARTA DO P. LEONARDO DO VALE POR...* [Paragraph | Section]
são e o mancebo vivo, e que vinha hum certo Indio com novas e mostras de ouro. Pello que, digo começarem jaa a terem algum começo de lume, pois, sendo tam faltos de saber e discrição humana, se lhes encaixou ser aquelle crucifixo alguma grande cousa, pois ningem o podia matar com espada nen **frecha**, e elle, sem se mover, matava a quem queria. E hé de crer que o christão não escapara de ser comido se o medo que tinhão à imagem de Christo Nosso Senhor hos não estorvara. {68.- BAÍA 26 DE JUNHO DE 1562 501- A00_0057.txt,N} Esta digressão de novas fiz
3. PADRE FERNÃO CARDIM. *III - INFORMAÇÃO DA MISSÃO DO...* [Paragraph | Section]
pau o abrem pelo meio, e acontece passarem um homem de parte a parte, e ir pregar no chão: exercitão-se de muito pequenos nestas armas, e são grandes frecheiros e tão certos que lhes não escapa passarinho por pequeno que seja, nem bicho do mato, e não tem mais que quererem meter uma **frecha** por um olho de um passaro, ou de um homem, ou darem em qualquer outra cousa, por pequena que seja, que o não fação muito ao seu alvo, e por isso são muito temidos, e tão intrepidos e ferozes que mete espanto. São como bichos do mato, porque entrão pelo sertão a caçar despídos e
4. PADRE FERNÃO CARDIM. *III - INFORMAÇÃO DA MISSÃO DO...* [Paragraph | Section]
que nenhum momento se apartão delle, quer vá pelas casas, quer para o mato, ou ande pelo terreiro, que para tudo tem liberdade, e commumente a guarda é uma que lhe dão por mulher, e também para lhe fazer de comer, o qual se seus senhores lhe não dão de comer, como é costume, toma um arco e **frecha** e atira á primeira galinha ou pato que vê, de quem quer que seja, e ninguém lhe vai á mão, e assim vai engordando, sem por isso perder o somno, nem o rir e folgar como os outros, e algunsandão tão contentes com haverem de ser comidos, que por nenhuma via consentirão ser resgatados
5. PADRE FERNÃO CARDIM. *III - INFORMAÇÃO DA MISSÃO DO...* [Paragraph | Section]

Find: word padding Next Previous Highlight all Match case Reached end of page, continued from top

Done

Busca no corpus (2)

File Edit View History Bookmarks Tools Help

http://moodle.icmc.usp.br/cgi-bin/newphilo/showrest_?kwic.6.1.19459.0.48.mil

Google BR-Linux.org | Linux ... dict.org Wikipedia Bandeijão

1. A00_0034 (bib:p.0)ovoão o Engenno sem yndio atirar **frecha**, antes se cree que, ja satisfeitos da mor
2. A00_0057 (bib:p.0)gem o podia matar com espada nen **frecha**, e elle, sem se mover, matava a quem queri
3. A00_0751 (bib:p.0) tem mais que quererem meter uma **frecha** por um olho de um passaro, ou de um homem,
4. A00_0751 (bib:p.0), como é costume, toma um arco e **frecha** e atira á primeira galinha ou pato que v
5. A00_0751 (bib:p.0) vivem de rapina e pela ponta de **frecha**, comem a mandioca crua sem lhes fazer mal,
6. A00_0729 (bib:p.0)am por elles, lhes deixam alguma **frecha**, ou pennas ou outra cousa como por offerta
7. A00_0749 (bib:p.0)aminas são tão duras que nenhuma **frecha** as pode passar se lhe não dá pelas ilhar
8. A00_0749 (bib:p.0)uns destes quando lhes tirão uma **frecha** a tomão na mão e tornão com ella atirar
9. A00_0749 (bib:p.0)ta, então se acolhem. Tomão-se á **frecha**, e assovião-lhe primeiro; tambem se tomã
10. A00_0749 (bib:p.0)rmado, e quando se armão não ha **frecha** que os passe; são muito pintados de varia
11. A00_0750 (bib:p.0)tem mais que quererem meter uma **frecha** por um olho de um passaro, ou de um homem,
12. A00_0750 (bib:p.0) como é costume, toma um arco e **frecha** e atira á primeira galinha ou pato que v
13. A00_0750 (bib:p.0)vivem de rapina e pela ponta de **frecha**, comem a mandioca crua sem lhes fazer mal,
14. A00_0064 (bib:p.0)omer a seu sñor peixe sempre de **frecha**.estes vi (e filo ja) botarem huã frecha p
15. A00_0064 (bib:p.0)stes vi (e filo ja) botarem huã **frecha** pera o ar e primeiro q'venha ao chaõ desp
16. A00_0064 (bib:p.0)zia faria qualquer cousa cõ huã **frecha** todo o brasil pa rece hu" jardim fresco p
17. A00_0064 (bib:p.0)go se afastaõ dali porq'tiraõ a **frecha** e tornam atirar cõ ella e saõ mui acerte
18. A00_0065 (bib:p.0) o q'l^o lhe toca na gerra ou o **frecha** desse ha de ser e o mesmo he da cassa porc
19. A00_0065 (bib:p.0)Rabo te" o mesmo e naõ os passa **frecha** Ha outros m^to diuerços tatupeba tatuquax
20. A00_0065 (bib:p.0) parte q'naõ mo rra logo tira a **frecha** e torna atirar cõ ella o bogio ao Jndio e
21. A00_0065 (bib:p.0)o Jndio atirãdolhe de Riba cõ a **frecha** e assi se afasta e escõde o Jndio logo co
22. A00_0065 (bib:p.0)s e assi o embedeã e depois a **frecha** e cõ fisgas Redes e a maõ os tomaõ e he
23. A00_1316 (bib:p.0) os Indios no caminho mataram á **frecha** um bogio grande de barba. Accudiram logo
24. A00_0700 (bib:p.0) todos eram 250, mais de arco e **frecha**, os Indios e os brancos de espada e rodell
25. A00_0157 (bib:p.0) peixe que se há-de matar com a **frecha**. Muitas cousas nos contam da sua fertilida
26. A00_0815 (bib:p.0)erio levemente no peito com uma **frecha** a um Francisco Carvalho de Lordelo e acodi
27. A00_0815 (bib:p.0)-nos, e servindo-nos sem arco e **frecha**, e admirando muito as nossas armas. Offere
28. A00_0815 (bib:p.0)xá, vive aldeiado, usa de arco, **frecha**, e porrete, é muito claro, e bem feito; a
29. A00_1241 (bib:p.0)tro mil Indios, uzaõ de arco, e **frecha**, sem mais outras armas, e nella naõ havia
30. A00_1133 (bib:p.0)hamada Cambajubaveya quer dizer **frecha** que abrio hum braço (porque o gentio atir
31. A00_0617 (bib:p.0). Dos Phenices uzarem de arco e **frecha** e paos tostados. A diversidade de lingoas
32. A00_0614 (bib:p.0)olho, porem Tabyra arrancando a **frecha** e com ella o olho; Brama como Touro, ruge
33. A00_0267 (bib:p.0) tambem alguns Indios de arco e **frecha**, cobertos todos da commodidade daquelles m
34. A00_0742 (bib:p.0)dio veio offerecer de venda uma **frecha** ou uma {A00_0742 revisado-- 206-.N} arma
35. A00_0742 (bib:p.0)ha de algodão em rama, com esta **frecha**, que chamam de esgravatana, se faz tiro, o
36. A00_0742 (bib:p.0)m sim alguma pesca, que fazem á **frecha**, ou talvez a linha; os costumes são os me
37. A00_0742 (bib:p.0)talvez n'esta altura com arco e **frecha** tres ou quatro indias disputarem o desemba
38. A00_0742 (bib:p.0)ccedido muitas vezes, usarem de **frecha** as mulheres no sertão: e a um criado noss

Done

Buscas

- Buscas bibliográficas
 - Todas as obras (busca em branco)
 - Autor, título, data, período
 - Usar maiúsculas
- Buscas avançadas
 - Colocações
 - Frequência por século

Exemplo de busca bibliográfica

The screenshot shows a web browser window titled "Philologic Results - Iceweasel". The address bar contains the URL "http://moodle.icmc.usp.br/cgi-bin/newphilo/search3t?dbname=miler". The search results are displayed as a list of entries, each starting with a checkbox and followed by the author's name, year, and a brief description of the document. The search term "word padding" is entered in the search box at the bottom, and the status bar shows "Done".

Philologic Results - Iceweasel

File Edit View History Bookmarks Tools Help

http://moodle.icmc.usp.br/cgi-bin/newphilo/search3t?dbname=miler Google

- LUIZ ARANHA DE VASCONCELLOS [1625], VII. - DIVERSOS DOCUMENTOS SOBRE O MARANHÃO E O PARÁ: 14. INFORMAÇÃO DE LUIZ ARANHA DE VASCONCELLOS SOBRE O DESCOBRIMENTO DO RIO DAS AMAZONAS. 1625. () [word count] [A00_0590].
- LUIZ AGOSTINHO VARELLA [1754], RELAÇAM, | E NOTICIA DE VARIOS SUCCESSOS | ACONTECIDOS | NO BRAZIL. | COPIA DE HUMA CARTA, QUE POR HUMA | DAS NÃOS QUE PROXIMAMENTE CHEGARAÕ MANDOU A HUM | SEU CORRESPONDENTE NESTA CORTE LUIZ AGOSTINHO VA- | RELLA ASSISTENTE NO RIO DE JANEIRO, COM OUTRAS | MAIS NOICIAS, EXTRAHIDAS DE VARIAS CARTAS | MAIS RECOPIADAS NESTA RELAÇÃO. | () [word count] [A00_1132].
- DOMINGOS PEREIRA AMARAL COUTINHO E LUIZ ANTONIO DE ARAUJO ALVES [1781], 08. AUTO DE MEDIÇÃO () [word count] [A00_1325].
- LUIZ DOS SANTOS VILHENA [1801], CARTA A D. RODRIGO DE SOUZA COUTINHO () [word count] [A00_0892].
- LUIZ DOS SANTOS VILHENA [1801], CARTA A D. RODRIGO DE SOUZA COUTINHO () [word count] [A00_0893].
- LUIZ DOS SANTOS VILHENA [1801], CARTA A PATRIFILO: () [word count] [A00_0848].
- LUIZ DOS SANTOS VILHENA [1801], CARTA DECIMA OITAVA: () [word count] [A00_0844].
- LUIZ DOS SANTOS VILHENA [1801], CARTA DECIMA SETIMA: () [word count] [A00_0843].
- LUIZ DOS SANTOS VILHENA [1801], CARTA DECIMA SEXTA () [word count] [A00_0841].
- LUIZ DOS SANTOS VILHENA [1801], CARTA DÉCIMA NONA: () [word count] [A00_0845].
- LUIZ DOS SANTOS VILHENA [1801], CARTA VIGESIMA QUARTA () [word count] [A00_0896].
- LUIZ DOS SANTOS VILHENA [1801], CARTA VIGESIMA SEGUNDA () [word count] [A00_0894].
- LUIZ DOS SANTOS VILHENA [1801], CARTA VIGESIMA TERCEIRA () [word count] [A00_0895].

Find: word padding Next Previous Highlight all Match case Reached end of page, continued from top

Done

Exemplo de colocações

Philologic Results - Iceweasel

File Edit View History Bookmarks Tools Help

http://moodle.icmc Google

Searching **Entire Database** for **frecha**.
Number of Unique Forms: 1

Search Terms: frecha

Your search found 49 occurrences.

Keywords found (with occurrences): frecha (49)

The 120 most common words are being filtered from this report. To include filtered words select "Turn Filter Off" on the search-form.

Ranking	Within 5 Words on Either Side	Ranking	Within 5 Words to Left only	Ranking	Within 5 Words to Right only
1	arco (12)	1	arco (11)	1	olho (3)
2	cõ (5)	2	matar (3)	2	ella (3)
3	olho (4)	3	toma (2)	3	cõ (3)
4	muitas (3)	4	rapina (2)	4	primeiro (2)
5	matar (3)	5	quererem (2)	5	primeira (2)
6	ella (3)	6	ponta (2)	6	muitas (2)
7	atirar (3)	7	meter (2)	7	mandioca (2)
8	tomão (2)	8	huã (2)	8	galinha (2)
9	toma (2)	9	dispararam (2)	9	crua (2)
10	rapina (2)	10	cõ (2)	10	cousas (2)
11	quererem (2)	11	costume (2)	11	comem (2)
12	primeiro (2)	12	alguma (2)	12	claro (2)
13	primeira (2)	13	yndio (1)	13	atirar (2)
14	ponta (2)	14	vive (1)	14	atira (2)
15	peixe (2)	15	viuẽ (1)	15	vi (1)
16	meter (2)	16	vezes (1)	16	vasio (1)
17	mandioca (2)	17	venda (1)	17	tostados (1)

Done

Philologic x Notas

Amanhã parto outra vez a Douvres a embarcar-me, e procurarei com toda a brevidade achar-me aos pés de V. Ex. ^a. Guarde Deus a V. Ex. ^a muitos anos como desejo. Londres, e Setembro 26 de 647. Vão também os maços das senhoras **Marqueza** e **Condessa**¹ e outro para o senhor Residente, a quem beijo a mão.

1 **Marquesa** de Niza e **Condessa** da Vidigueira.

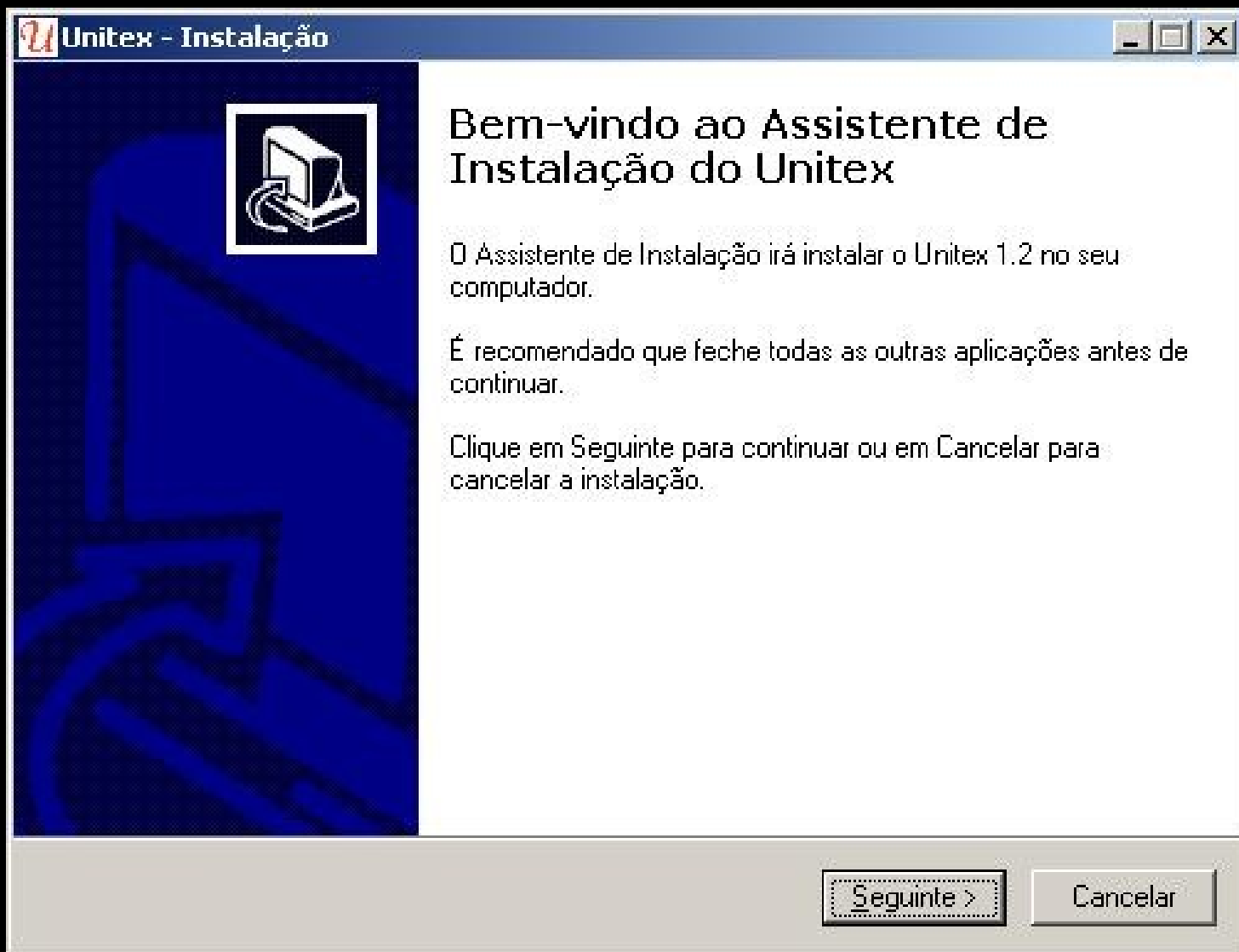
Unitex-milenio

- Adaptado às necessidades do Projeto do Dicionário Histórico:
 - Contém o corpus digitalizado
 - Suporta símbolos do Português Histórico (como o S-longo)
 - Léxicos: contemporâneo, abreviaturas, variações ortográficas
 - Ferramenta para busca nos léxicos

Instalando o Unites

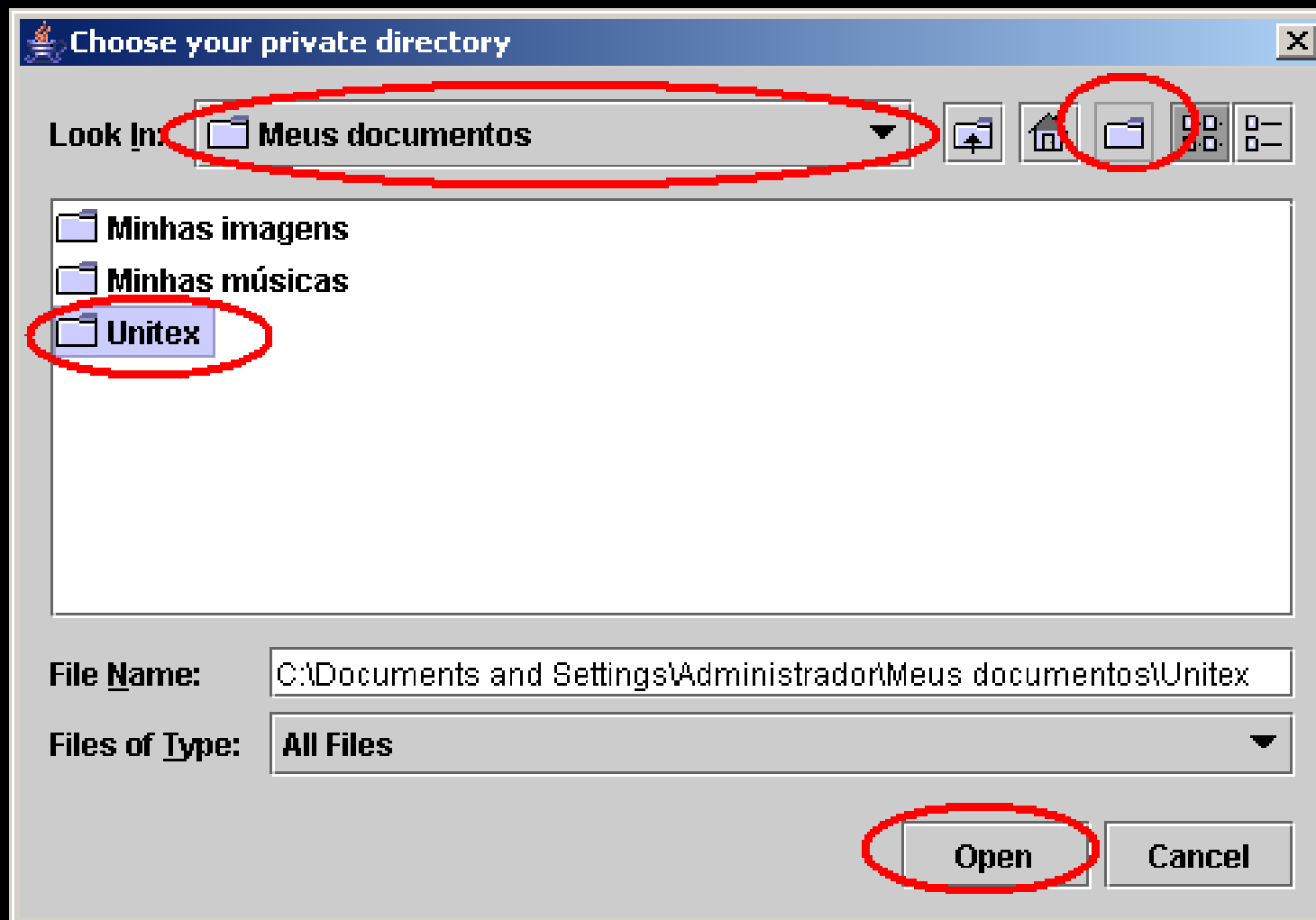
- Diretamente pelo CD ou pela Internet:
 - <http://moodle.icmc.usp.br/milenio/>
- Passos:
 - 1 - Instalar o Java
 - 2 - Instalar o Unites (versão 2 recomendável)
 - 3 - Instalar o Dicionário
- Instalação do tipo “Prosseguir, Prosseguir, Finalizar”

Instalando o Unitex (2)



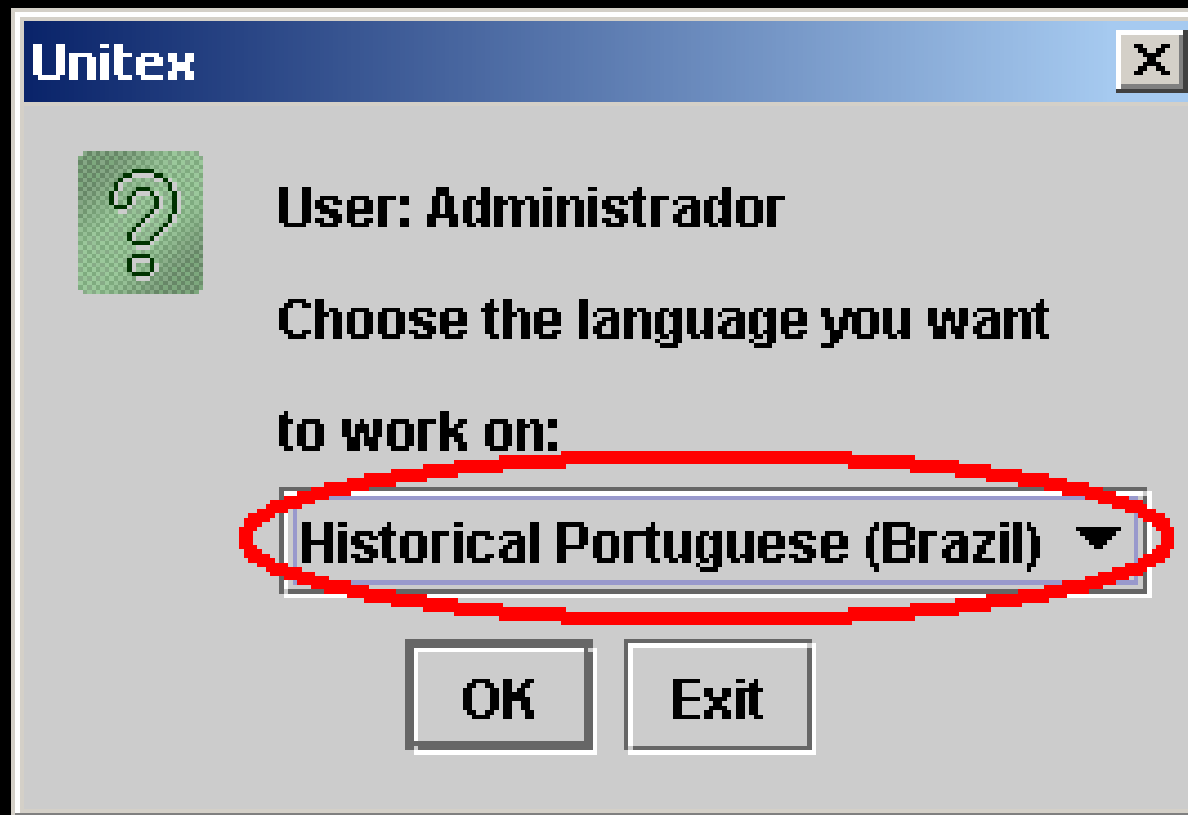
Abrindo o Unitex pela primeira vez

- Escolher um diretório de trabalho (dica: Unitex em Meus Documentos)



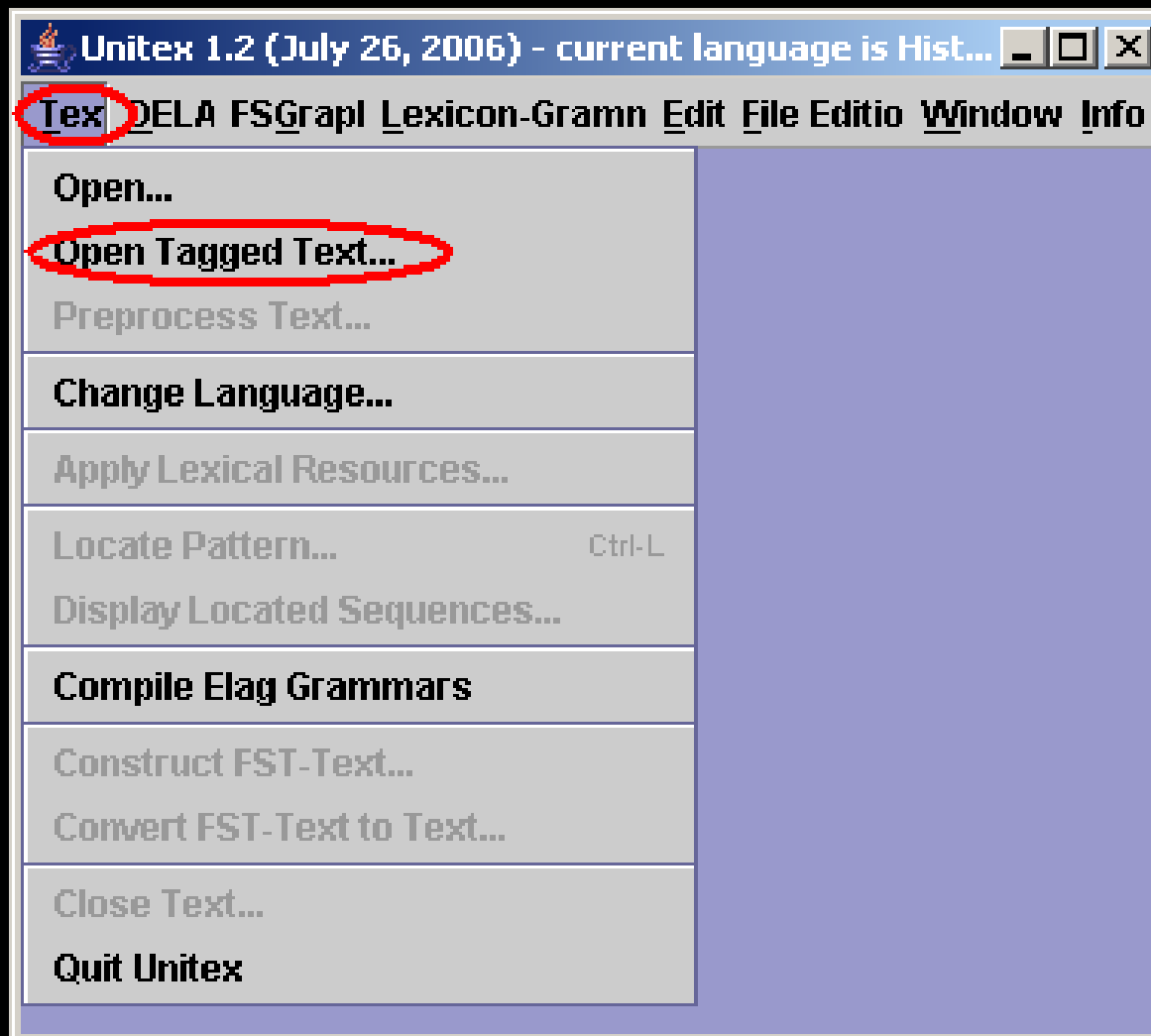
Abrindo o Unitex pela primeira vez (2)

- Escolher o idioma “Historical Portuguese (Brazil)”



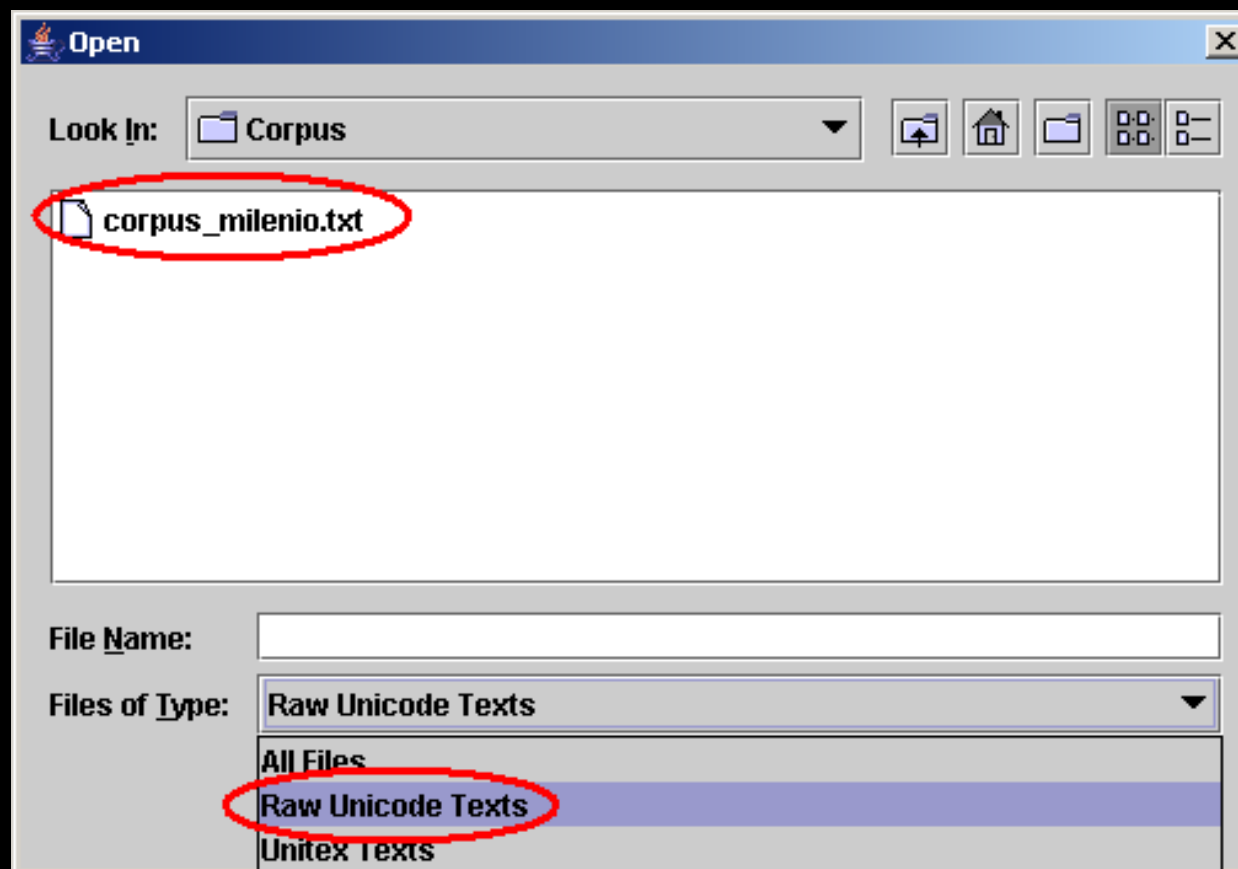
Processando o corpus

- File -> Open Tagged Text



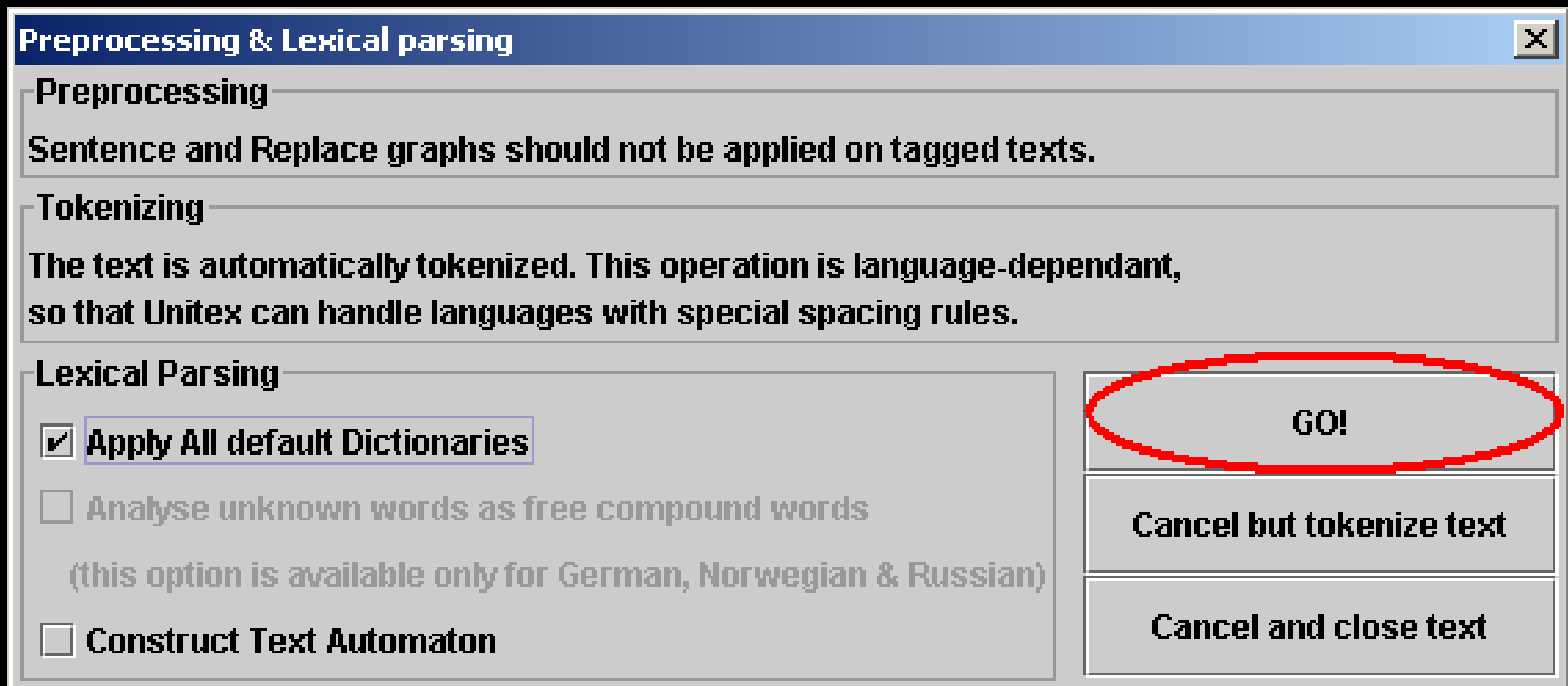
Processando o corpus (2)

- Unitex Texts -> “Raw Unicode Texts” -> “corpus_milenio.txt” -> “Open”



Processando o corpus (3)

- Clicar em “Yes” e depois em “GO”



Córpus aberto

The screenshot displays the Unitex 1.2 software interface. The main window shows a corpus file path: **C:\Documents and Settings\Administrador\Meus documentos\Historical Port...**. Below the path, it indicates: **0 sentence delimiter, 4973890 (104459 diff) tokens, 2281784 (96752) simple forms, 62165 ...**. The main text area contains a corpus entry for a historical document, including a title and several paragraphs of text in Portuguese.

Overlaid on the main window are two smaller windows:

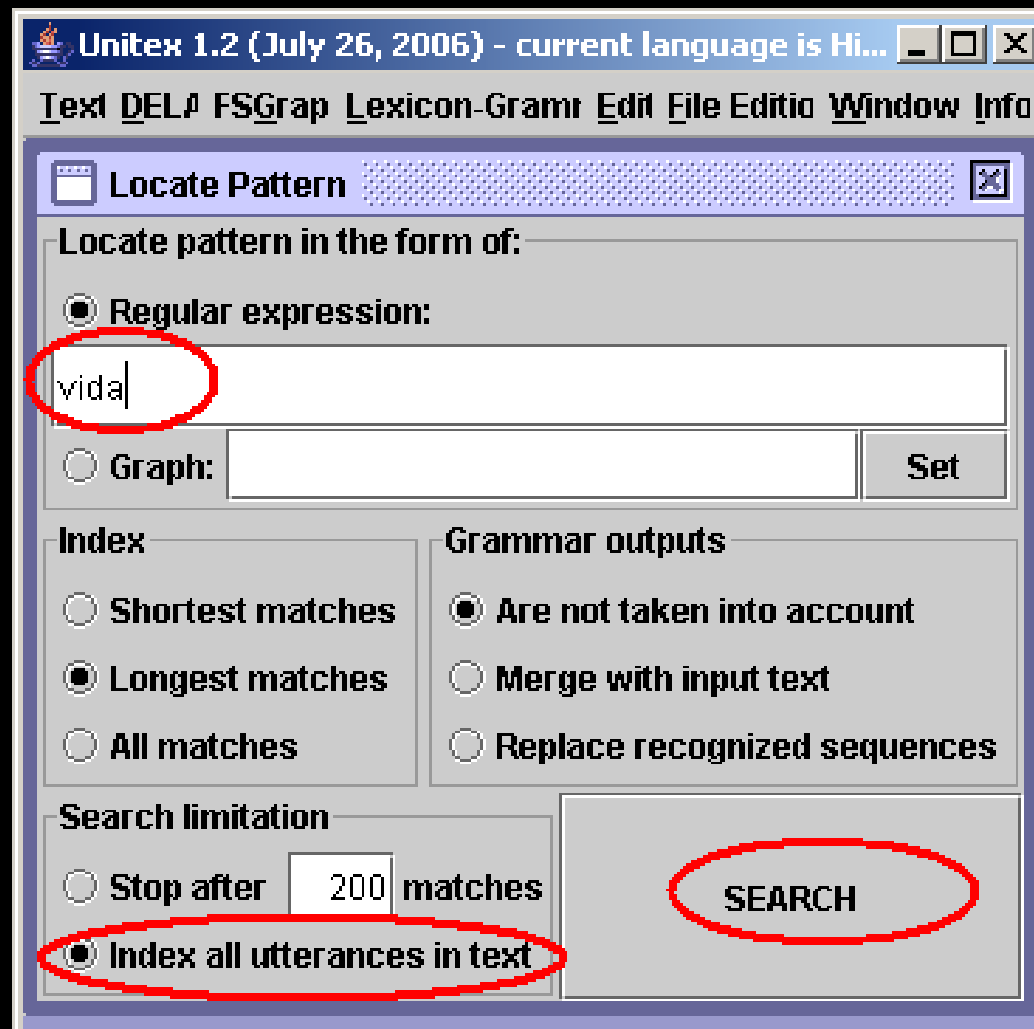
- Token list**: A window showing a list of tokens sorted by frequency. The top entries are: 2281145 (space), 181992 (comma), 101935 (de), 96850 (e), 90186 (period), 87997 (que), 62724 (a), 53374 (o), and 39258 (se).
- Word Lists in C:\Documents and Settings...**: A window showing two lists of words. The left list is titled **DLF: 49124 simple-word l...** and the right list is titled **ERR: 51850 unknown sim...**. The left list includes morphological tags like **@, anos. N+ABREV:ms**, **a, .ABREV:ms**, **a, .N:ms**, **a, .PREP**, **A, alteza. N+ABREV:fs**, **A, alvará. N+ABREV:ms**, **A, Amaro. Npr+ABREV:r**, and **A, Ana. Npr+ABREV:fs**. The right list includes words like **entrancia**, **entranhámos**, **entranhavão**, **entranhavel**, **entranhavelmente**, **entranhàvelmente**, **entranhos**, **Entrão**, **entraõ**, **entrão**, **entrará**, **entrára**, **entrarão**, **ENTRÁRÃO**, **Entrarão**, **entrarão**, **entrariamos**, **Entraron**, and **entrarse**.

Abrindo o córpus depois de processado

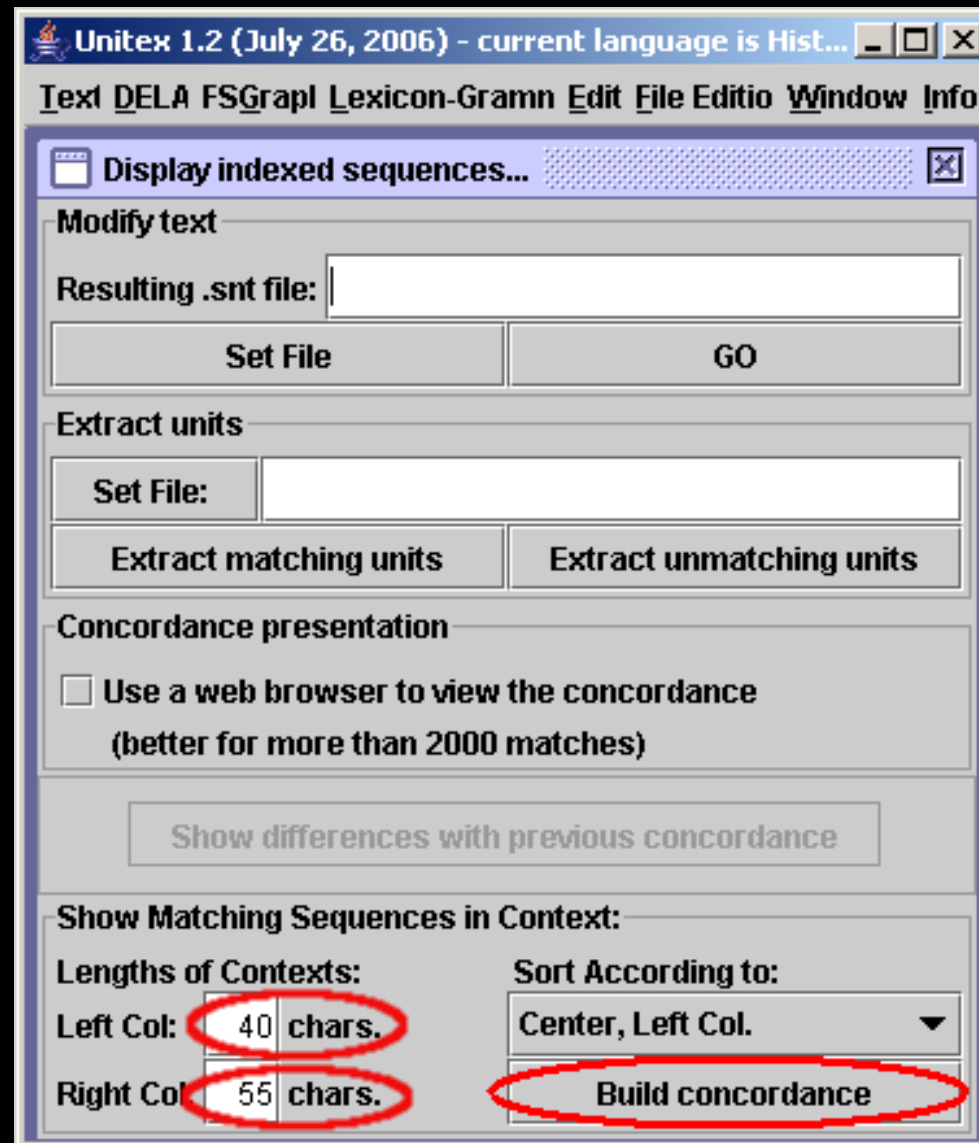
- Clicar em “File” e “Open Tagged”
- Clicar em “corpus_milenio.snt” e em “Open”

Buscando

- Text -> Locate Pattern



Personalizando a busca



Resultados da busca

The screenshot shows the Unitesx 1.2 software interface. The title bar reads "Unitesx 1.2 (July 26, 2006) - current language is Historical Portuguese (Brazil)". The menu bar includes "Text", "DELA", "FSGraph", "Lexicon-Grammar", "Edit", "File Edition", "Windows", and "Info". The main window title is "Concordance: C:\Documents and Settings\Administrador\Meus documentos\Historical Portuguese". Below the title bar, it indicates "1044 matches" and has a checked "Enable links" option. The main content area displays a concordance of the word "vida" in Historical Portuguese, with the word highlighted in blue in the original image. The text is a long paragraph from a historical document, discussing the language and customs of the Tupinambás and the word "vida" in various contexts.

e que fallámos, o qual tem a linguagem, [vida](#) e costumes e gentilidades dos Tupinambás, ainda qu
gente do Brazil semelhante no parecer, [vida](#) e costumes dos Tupinambás, e na linguagem não tem
a outra, e tem as mesmas gentilidades, [vida](#) e costumes, e são amigos uns dos outros. São estes
r grandes informações de suas divisões, [vida](#) e costumes: mas pois ao presente não é possível, t
dias Santos, cujos são os Vasconcellos, [Vida](#) do P. Anchieta, liv. II, cap. I. {Tipologias - Mar
de navegar. Padre Manoel da Fonseca. - [Vida](#) do Padre Belchior de Pontes. Pedro Taques de Almei
ecedentes das cousas do Brasil. » » » - [Vida](#) do Padre João de Almeida. » » » - » » » José de An
ta é summariamente a primeira parte da. [vida](#) do Padre João da Souto Maior, a qual junta com a q
ticos, moraes, e panegyricos. 4. 4 Tom. [Vida](#) do excellentissimo e reverendissimo D. Frey Jozê F
I Alvares Sollano anno 1754, 4. Compoz. [Vida](#) do veneravel padre Jozé de Anchieta escrita na lin
nes virtudes gloriosa fama. CAPITULO 10 [VIDA](#) E PRECIOSA MORTE DA PENITENTE JOANNA DE JE
inada de Jesus, o seu livro intitulado: [Vida](#) do Padre Belxior de Pontes, impresso em Lisboa no
s quaes as horas se passam vagarosas, a [vida](#) fastidiosa e pensativa, e sem os quaes nem ha riqu
com outros dizendo: «Não tem F. e F. a [vida](#) boa como nós; e se algum'ora a vierem a {66. - BAÍ
de muito longe possa tocar ao decoro. A [vida](#) está só na mão de Deus, e esta é a ocasião em que
endeo, que perdendo EIRey D. João IV. a [vida](#) as mãos de hum traidor, seria o meyo de assegurar
Aras de Marte em holocausto o sangue; a [vida](#) em sacrificio pela liberdade da Patria. {A00_0824.
icioso logro pelo qual perdeu a Irmã a [vida](#) e elle a honra. . Na aurora da vida, na primavera
im a morte nos assalteia, e nos rouba a [vida](#), sem sabermos o dia, nem a hora : Quia nescitis di
não há descansar, e nisto se gasta cá a [vida](#) dos nossos, com que se têm ganhado em todo o Brasi
a e costumes d'estas mulheres. Começa a [vida](#) e costumes dos Tapuias. Como a tenção com que nos
smorecidos tinham perdido a esperança á [vida](#); mas a divina providencia, que por caminhos sempre
do, em que estava, e levaria ariscada a [vida](#) a hum veneno, ou exposta a continuos dissabores. S
morte tão sentida, como era estimada a [vida](#) de Bento do Amaral da Silva; porém morreu quem mor
{A00_0824 - 172,.N} sempre sepultada a [vida](#), e o nome. O olandez que se vio descuberto, para q
do mestre, e depois capitão. Deixada a [vida](#) militar seguio {A00_0822 - 41,.N} a ecclesiastica
duram na terra, sem se seccarem, toda a [vida](#); se andam limpos de herva, que entre elles nasce,
dos faz a oração funebre. Relata toda a [vida](#) do defunto, as suas generosas façanhas e açções :

Problemas com Unites 1.2

- Links não funcionam adequadamente
- Solução 1:
 - Modificar a quantidade caracteres buscada pelo concordanceador.
 - Buscas simples: 150 caracteres
 - Concordâncias: 1500 caracteres
- Solução 2:
 - **Unites 2.0** (versão em desenvolvimento)

Referenciando o córpus

{A00_0568 revisado - LIVRO SEGUNDO 89,.N}

Senhora da Conceição ; principia a sua jurisdicção no estreito de Santos, onde está a Villa deste nome, rodeada do mar, e tres legoas por costa distante da de S. Vicente ; ficando as duas (...) As que se **comprehendem** nas jurisdicções de ambas as Capitanias para a parte do mar, são a de S. Sebastião, em cujo (...). As Villas da Serra acima são, Jacarahi, Penhamunhangaba, Goaratinguitâ, Corutubâ, Sorocaba, Utû, Jundiahi, Paranaíba, Taubatê, Mongi, onde ha hum

{A00_0568 revisado - 90 AMERICA PORTUGUEZA,.N}

e dos marmellos as mais finas marmelladas, e já de presente excellente jalea. Abunda de muitos generos de mimosas carnes, e caças gostossissimas ; cultiva no seu termo muitas quintas deliciosas (...)

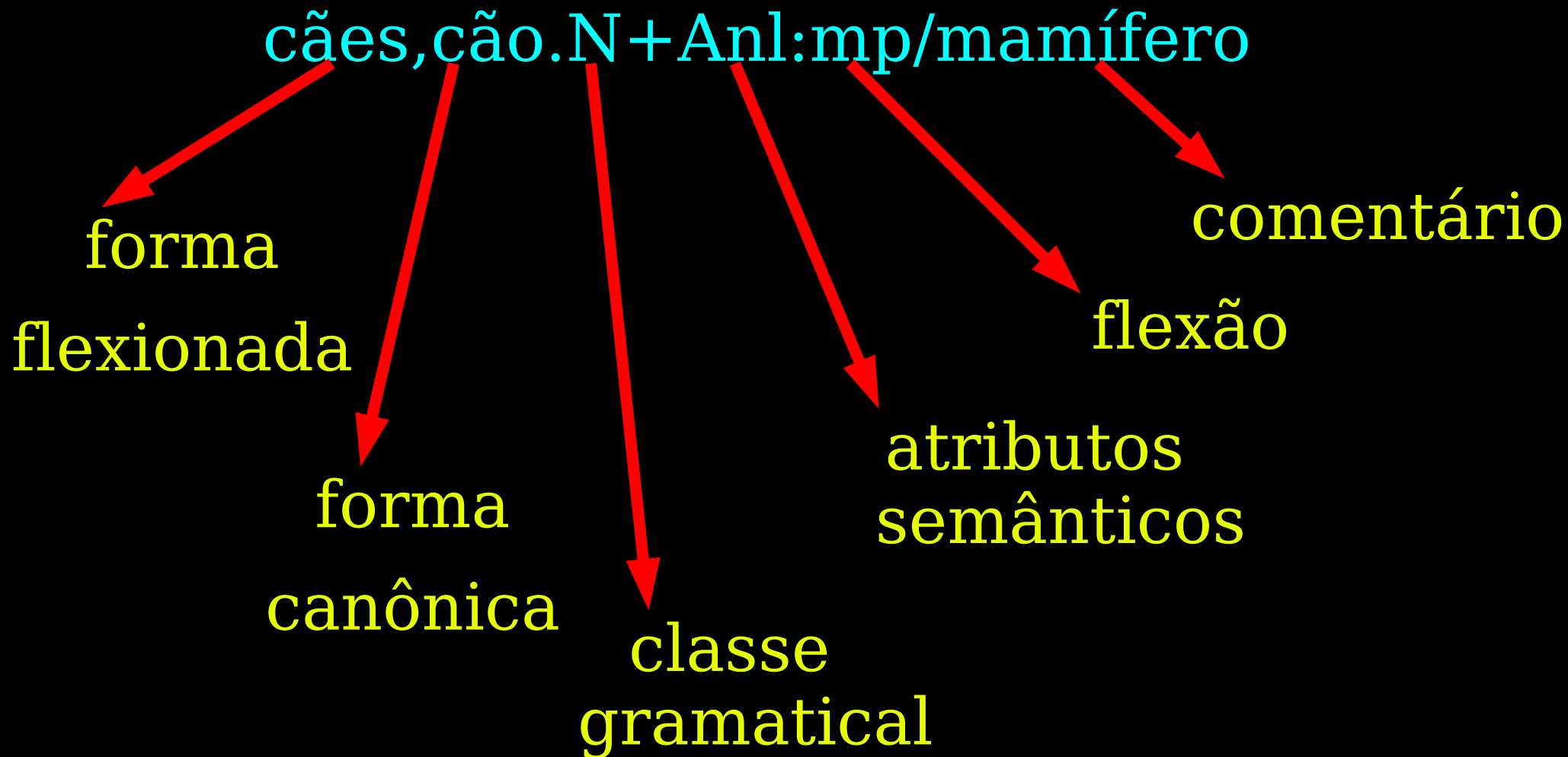
Proposta para referências

- Código do arquivo + número da página
- Referência criada
 - [A00_0568, 89]
- Autor, data, título e outra informações são obtidas automaticamente
 - Formato ABNT pode ser gerado

Exemplos de buscas (2)

- Expressões de busca:
 - Palavras em minúsculas: <MIN>
 - Palavras em maiúsculas: <MAJ>
 - Iniciadas em maiúsculas: <PRE>
 - Palavras que não pertencem ao dicionário: !<DIC>
- Buscas combinadas: sr. <PRE>
- Dica: observe o uso de maiúsculas e minúsculas

Léxicos no formato DELA



Códigos gramaticais

Código	Classe
A	Adjetivo
N	Substantivo
Prep	Preposição
etc	...

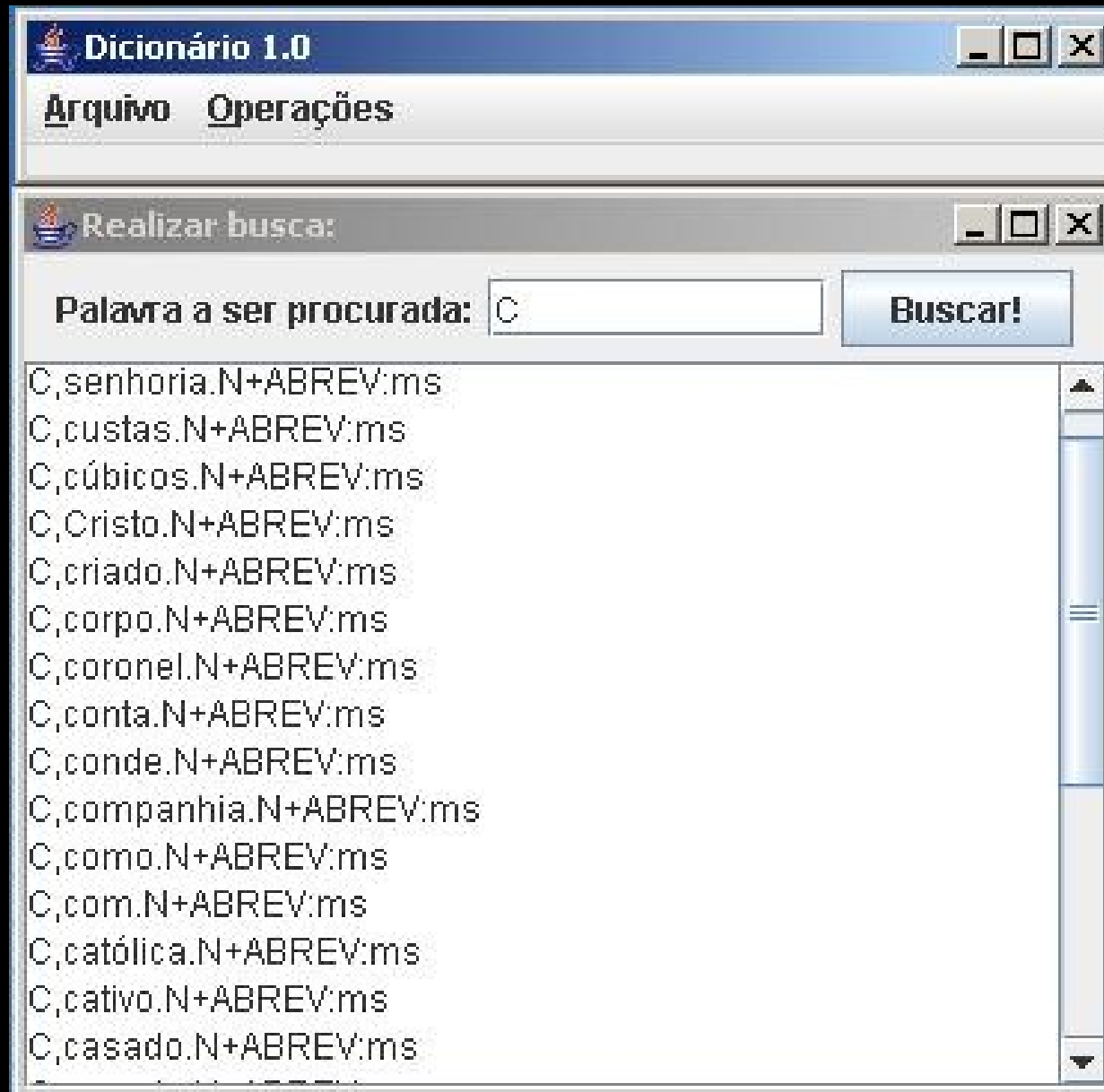
* A lista com os códigos mais usuais se encontra no manual

Códigos flexionais

Código	Flexão
FP	Feminino plural
P3s	3a. pessoa do singular
W	Infinitivo
etc	...

* A lista com os códigos mais usuais se encontra no manual

O programa Dicionário



Léxico de Português Contemporâneo

- Desenvolvido no mestrado de Marcelo Muniz (2005)
- ~ 850 mil entradas
- Exemplos de buscas:
 - <ser.N>: ser, seres
 - <ser.V>: ser, sou, és, foi, sendo, etc
 - <ser>: ser, seres, sou, és, sendo, etc
 - Sujeito a ambigüidade

Léxico de Abreviaturas Históricas

- Fontes
 - Flexor: digitalização pela Clarissa
 - Inventários e testamento: levantamento feito pela Livia
- Atualmente
 - 5.093 abreviaturas
 - 3.558 formas expandidas
 - Inventário + flexor A, B e C
 - D até Z em andamento

Léxico de Abreviaturas Históricas (2)

- 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

Léxico de Abreviaturas Históricas (3)

- Exemplos de buscas:
- <janeiro.ABREV>: Jan, Jan.^{ro}, Jan^{ro}, Janr.^o, Jan.^o, Jn^{ro}, Janr^o
 - Nota: Jan.^{ro} --> Jan.^{ro}
- <ABREV>: retorna todas as abreviaturas

Léxico de variações ortográficas

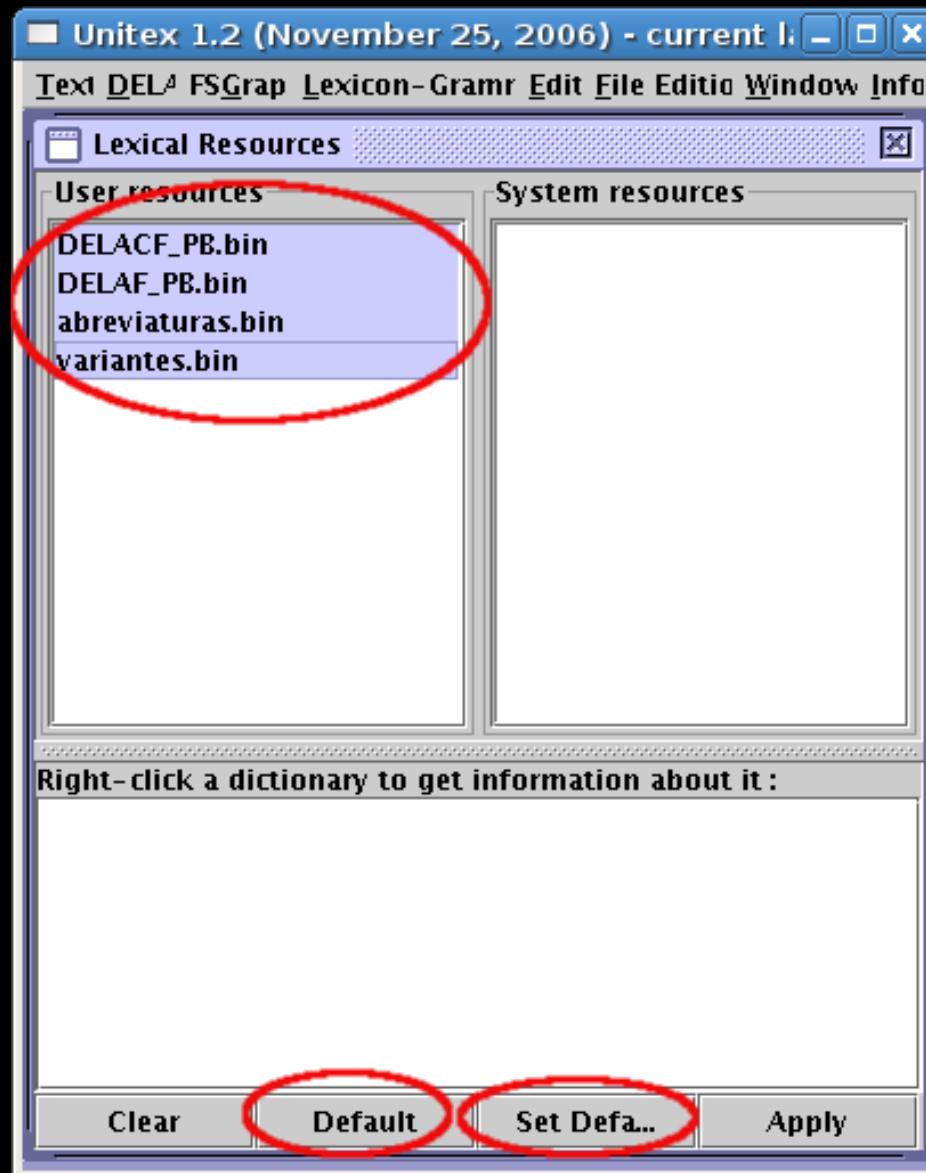
- Trabalho desenvolvido por Rafael Giusti
- Agrupamento automático de formas ortográficas correlatas
 - Ex.: muito, mui, mujto
- 10.859 agrupamentos
- 24.114 variantes
- Revisão manual necessária devido a pequenos erros de agrupamento

Léxico de variações ortográficas (2)

- Atualmente
 - Busca por forma normatizada
 - <foi.VAR>: foi, ffoi, foy, ffoy
- Futuramente
 - Busca por forma canônica
 - <ir.VAR>: foi, ffoi, foy, ffoy, vamos, vão, vão, vam, etc

Ativando os léxicos

- Text -> Apply Lexical Resources



Exemplos de buscas

- Palavras simples
 - Unitex: cabedal
 - Philologic: cabedal
 - Resultado: cabedal

Exemplos de buscas (2)

- Seqüências de palavras
 - **Unitex**: sua magestade
 - **Philologic**: sua magestade
 - **Resultado**: sua magestade

Exemplos de buscas ⁽³⁾

- União de palavras (“ou lógico”)
 - **Unitex**: capitão+capitão
 - **Philologic**: capitão|capitão
 - **Resultados**: capitão, capitão

Exemplos de buscas ⁽⁴⁾

- União de seqüências de palavras (ou lógico)
 - **Unitex**: (capitão mór)+(capitaõ mór)+(capitão mor)+(capitaõ mor)
 - **Philologic**: capitão|capitaõ mór|mór
 - **Resultados**: capitão mór, capitaõ mór, capitão mor, capitaõ mor

Exemplos de buscas (5)

- Infixos
 - Unitex: <<ct>>
 - Philologic: .*ct.*
 - Resultados: acto, benedictus, conjecturar, delicto, ...

Exemplos de buscas (6)

- Prefixos
 - Unitex: <<^san>>
 - Philologic: san.*
 - Resultados: Santiago, Santa, Santos, Santo, Santarem, ...
- Unitex 1.2 apresenta problemas na busca por prefixos.

Exemplos de buscas (7)

- Sufixos
 - Unitex: <<mente\$>>
 - Philologic: .*mente
 - Resultados: abastadamente, amigavelmente, asperamente, brevemente, ...

Exemplos de buscas ⁽⁸⁾

- Coringas para letras (wildcards)
 - Unitex: <<^m.l\$>>
 - Philologic: m.l
 - Resultados: mal, mel, mil

Exemplos de buscas ⁽⁹⁾

- Coringas para palavras
 - **Unitex**: gente <MOT> terra
 - **Philologic**: gente terra
 - Ativar opção “frase separada por 1 palavra”
 - **Resultados**: gente da terra, gente desta terra, gente em terra, ...

Exemplos de buscas ⁽¹⁰⁾

- Buscas exclusivas do Unitex
 - Variações de grafia: <foi.VAR>
 - Abreviaturas: <janeiro.ABREV>
 - Léxico contemporâneo: <ser.V>

Exemplos de buscas (11)

- Buscas exclusivas do **Philologic**
 - Listagem bibliográfica
 - Autor, título, data e período
 - Colocações
 - Frequência por século

Unitex x Philologic

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

Símbolos especiais - Acentos

Símbolo	Unicode	Exemplo
^	0302	quar [^] y (*)
~	0303	com [~] ande
-	0304	caca ⁻ o
¨	0308	mu [¨] y
ˆ	0309	s ^ˆ o ^ˆ mente
°	030A	(Å) Afonso
'	0313	tinha ['] o

(*) nome indígena

Símbolos especiais - Letras

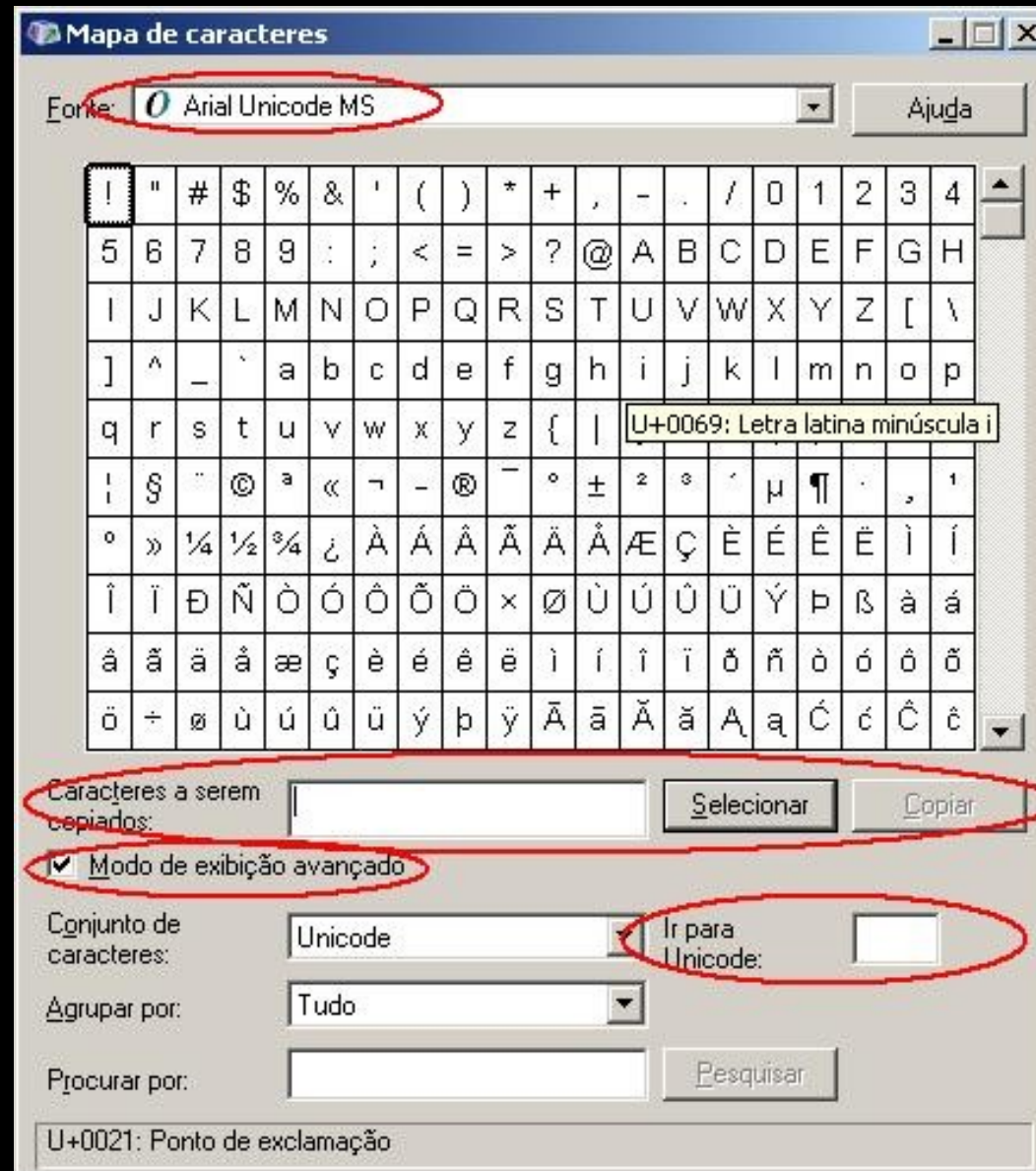
Símbolo	Unicode	Exemplo
Æ	00C6	Æthyopia (*)
æ	00E6	gratiæ (*)
œ	0153	coeteris (*)
§	00A7	parágrafo
¶	2132	¶ixit (*)
f	017F	difcurfo
f	0192	feito
e	0250	passade

(*) nomes em latin

Mapa de caracteres

- Para tratamento de símbolos que não podem ser digitados via teclado
- Pode ser utilizado com **Unitex** e **Philologic**

Mapa de caracteres



Proposta de ambiente Web

- O melhor dos dois mundos
 - As buscas e léxicos do Unitex
 - A centralização e simplicidade do Philologic
- Editor de verbetes agregado
- Enquanto o sistema é desenvolvido
 - Philologic é recomendado

Conclusões

- O **Unitex** fornece buscas poderosas e acesso a léxicos
- O **Philologic** tem como ponto forte a facilidade de uso, centralização e suporte a texto anotado
- O ambiente proposto agrega as funcionalidades de ambas as ferramentas e possui um editor de verbetes