Corpus para pesquisa linguística: compilação, gerenciamento e manipulação por meio de ferramentas computacionais

Gladis Maria de Barcellos Almeida
Departamento de Letras
Universidade Federal de São Carlos

Corpus/corpora

A utilização de corpus sempre foi um recurso empregado em estudos que tratam da língua/linguagem, o que mudou foi a concepção de corpus...

Uso de corpora em dicionários antigos (séculos XVIII e XIX)

- **Vocabulário Portuguez e Latino**

Elaborado pelo Pe. Rafael Bluteau e publicado entre 1712-1728 em 8 volumes.

Foi o primeiro dicionário para o qual foi fixado um corpus (Murakawa, 2001).

Uso de corpora em dicionários antigos (séculos XVIII e XIX)

- **Vocabulário Portuguez e Latino**

Elaborado pelo Pe. Rafael Bluteau e publicado entre 1712-1[illegible]mes.

Foi o prim[illegible]ara o qual foi fixado um corpus (Murakawa, 2001).

cerca de 406 obras contendo autores dos séculos XV a XVII

Uso de *Corpora* em dicionários antigos (séculos XVIII e XIX)

- **Vocabulário Portuguez e Latino**

> O corpus servia como fonte de exemplário de uso linguístico para as palavras que constavam da nomenclatura do dicionário.
>
> (Murakawa, 2001; 2006)

Uso de corpora em dicionários antigos (séculos XVIII e XIX)

- **Diccionario da Lingua Portugueza**, 2ª edição, de António de Morais Silva, publicado em 1813, o qual também se valeu de um corpus (Murakawa, 2006).

Afinal, o que é corpus?

- Concepção de corpus na Linguística
- Concepção de corpus na Linguística de Corpus

Para a Linguística

- **Segundo o *Dicionário de didáctica das línguas*, de Galisson & Coste (1983):**

"um conjunto finito de enunciados tomados como objeto de análise. Mais precisamente, conjunto finito de enunciados considerados característicos do tipo de língua a estudar, reunidos para servirem de base à descrição e, eventualmente, à elaboração de um modelo explicativo dessa língua. →

Para a Linguística

- ***Dicionário de didáctica das línguas*, de Galisson & Coste (1983):**

Trata-se, pois, de uma colecção de documentos quer orais (gravados ou transcritos) quer escritos, quer orais e escritos, de acordo com o tipo de investigação pretendido. As dimensões do corpus variam segundo os objectivos do investigador e o volume dos enunciados considerados como característicos do fenômeno a estudar. →

Para a Linguística

- ***Dicionário de didáctica das línguas,* de Galisson & Coste (1983):**

Um corpus é chamado exaustivo quando compreende todos os enunciados característicos. E é chamado selectivo quando compreende apenas uma parte desses enunciados" (p.169)

Para a Linguística

- **Segundo o *Dicionário de linguística*, de Dubois *et al.* (1993)**

Conjunto de enunciados a partir do qual se estabelece a gramática descritiva de uma língua. O corpus "não pode ser considerado como constituindo a língua, mas somente como uma amostra da língua. →

Para a Linguística

- ***Dicionário de linguística*, de Dubois et al. (1993)**

O corpus deve ser representativo, isto é, deve ilustrar toda a gama das características estruturais. Poder-se-ia pensar que as dificuldades serão levantadas se um corpus for exaustivo (...). →

Para a Linguística

- ***Dicionário de linguística*, de Dubois et al. (1993)**

Na realidade, sendo indefinido o número de enunciados possíveis, não há exaustividade verdadeira e, além disso, grandes quantidades de dados inúteis só podem complicar a pesquisa, tornando-a pesada. O linguista deve, pois, procurar obter um corpus realmente significativo. →

Para a Linguística

- ***Dicionário de linguística*, de Dubois et al. (1993)**

Enfim, o linguista deve desconfiar de tudo o que pode tornar o seu corpus não-representativo (método de pesquisa escolhido, anomalia que constitui a intrusão do linguista, preconceito sobre a língua)." (158-159)

Para a Linguística

- **Segundo o *Dicionário enciclopédico das ciências da linguagem*, de Ducrot & Todorov (2001)**

"...conjunto, tão variado quanto possível, de enunciados efetivamente emitidos por usuários da referida língua em determinada época" (p. 42)

Para a Linguística

- **Segundo o *Dicionário de Linguagem e Linguística*, de Trask (2004):**

"conjunto de textos escritos ou falados numa língua, disponível para análise" (p. 68) No mesmo verbete, o autor apresenta as vantagens de se utilizar corpus para a descrição da língua e sugere formas de armazenamento.

Para a Linguística de Corpus

"A corpus is a collection of pieces of language text in **electronic form**, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research." (Sinclair, 2005) [grifo nosso]

Para a Linguística de Corpus

"...conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, **dispostos de tal modo que possam ser processados por computador**, com a finalidade de propiciar resultados vários e úteis para a descrição e análise" (SANCHEZ, 1995, pp. 8-9, *apud* BERBER SARDINHA, 2000)

Para a Linguística de Corpus

Para outros dois eminentes linguistas de corpus,

> o emprego do termo corpus implica em conotações bastante específicas.

Segundo McEnery e Wilson (1996), a moderna

> noção de corpus carrega consigo pelo menos quatro características fundamentais:

Para a Linguística de Corpus

- ***amostragem e representatividade (sampling and representativeness):*** um corpus deve ter uma amostragem suficiente da língua ou variedade de língua que se quer analisar para obter-se o máximo de representatividade desta mesma língua ou variedade de língua;

Para a Linguística de Corpus

- ***tamanho finito (finite size):*** com exceção de corpus-monitor, todo corpus tem um tamanho finito, por exemplo: 500 mil palavras, 1 milhão de palavras, 10 milhões de palavras, etc;

Para a Linguística de Corpus

- ***tamanho finito (finite size):*** com exceção de *corpus*-monitor, todo *corpus* tem um tamanho finito, por exemplo: 500 mil palavras, 1 milhão de palavras, 10 milhões de

Corpus-monitor é aquele que pode receber novos textos e tornar-se cada vez maior. É um corpus útil para Lexicografia, por exemplo, já que é necessário observar palavras novas na língua ou palavras já conhecidas mas com emprego diferente.

Para a Linguística de Corpus

- ***formato eletrônico (machine-readable form):*** segundo McEnery e Wilson (1996), atualmente o emprego do termo corpus significa admitir necessariamente que os textos estejam no formato eletrônico.

Para a Linguística de Corpus

- ***formato eletrônico (machine-readable form):*** segundo McEnery e Wilson (1996), atualmente o significa adn textos esteja

Vantagens:

- os corpora podem ser pesquisados e manipulados de forma mais rápida;
- os corpora podem ser mais facilmente enriquecidos com informação extra.

Para a Linguística de Corpus

- ***referência padrão (standard reference):*** de acordo com McEnery e Wilson (1996), existe um entendimento tácito de que um corpus constitui uma referência padrão para a variedade de língua que ele representa, pressupondo que o corpus esteja disponível para outros pesquisadores, em outras palavras, é o que se tem chamado de **reuso** do corpus.

Para a Linguística de Corpus

- ***referência padrão (sta*** ordo com McEnery e Wilson entendimento tácito ui uma referência padrã ngua que ele representa, pr esteja disponível para outros pesquisad em outras palavras, é o que se tem chamado de **reuso** do corpus.

- diferença marcante entre a concepção de corpus para a Linguística e para a Linguística de Corpus
- característica inerente ao corpus

Linguística e Linguística de Corpus: 2 grandes diferenças

- o formato computadorizado do corpus e
- a sua posterior disponibilização para outras pesquisas

Web como corpus

Se a Linguística de Corpus descarta livros, revistas e outros textos impressos considerados corpus pela Linguística, ela também descarta a Web como corpus, ainda que os textos estejam disponíveis e em formato eletrônico, pelo fato de suas dimensões serem desconhecidas, estar continuamente mudando e pelo fato de não ter sido projetada a partir de uma perspectiva linguística.

Web como corpus

Entretanto, é a própria Web que vai facilitar a distribuição e o livre acesso de corpora criados em vários projetos, reforçando uma das características de corpus citadas por McEnery e Wilson (1996).

Web como corpus

Entretanto, é a própria Web que vai facilitar a distribuição e o livre acesso de corpora criados em vários projetos, reforçando uma das características de corpus citadas por McEnery e Wilson (1996).

Vale assinalar que há autores que consideram a Web um corpus, é o caso de Kilgarriff e Grefenstette (2003).

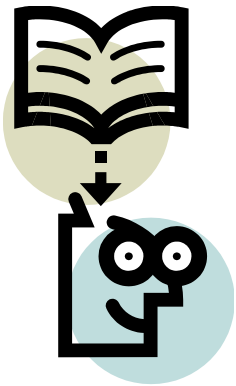
O computador

O surgimento do computador interfere diretamente não só na concepção que se tem de corpus como também na sua forma de armazenamento e exploração, o que acaba afetando os resultados de análise...

→

O computador

...já que os recursos oferecidos pelo computador permitem que uma quantidade antes inimaginável de textos possa ser processada na tela em questão de segundos, fazendo com que muitas hipóteses sobre determinados fenômenos linguísticos possam ser testadas rápida e eficientemente.



Contribuições

computador

Contribuições

computador

Linguística computacional

Contribuições

computador

ferramentas computacionais voltadas para PLN do português (Br)

Linguística computacional

Corpus na pesquisa linguística

Nesse sentido, a pesquisa descritiva volta a ter um amplo desenvolvimento, pois a possibilidade de lidar com grandes corpora permite a observação e descrição de fenômenos linguísticos recorrentes antes impossível de perceber, dado que os procedimentos de observação e descrição contavam apenas com recursos manuais.

E a Linguística de Corpus?

Abordagem que se ocupa "da coleta e da exploração de corpora, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por computador." (BERBER SARDINHA, 2004)

Requisitos para a elaboração de um *corpus* computadorizado

1. Os textos devem ser ***autênticos***. Por autenticidade, compreende-se:

- os textos devem ter sido escritos em linguagem natural, não podendo ser textos "produzidos com o propósito de serem alvo de pesquisa linguística";
- os textos devem ser escritos por falantes nativos, exceto se se tratar de *corpora* de aprendizes, aqueles *corpora* cujos textos são provenientes de falantes que estão aprendendo uma língua estrangeira.

(BERBER SARDINHA, 2000)

Requisitos para a elaboração de um corpus computadorizado

1. Os textos devem ser ***autênticos***. Por autenticidade, compreende-se:

- os textos devem ter sido escritos em linguagem natural, não podendo ser textos "produzidos com o propósito de serem alvo de pesquisa linguística";
- os textos devem ser escritos por falantes nativos, exceto se se tratar de corpora de aprendizes, aqueles *corpora* cujos textos são provenientes de falantes que estão aprendendo uma língua estrangeira.

(BERBER SARDINHA, 2000)

Requisitos para a elaboração de um corpus computadorizado

2. O corpus deve ter *representatividade*, isto é, ser representativo da língua ou de uma variedade de língua que ser deseja pesquisar. Idealmente, um corpus deve ser elaborado de forma a representar determinadas características linguísticas da comunidade cuja língua está sob análise (Sinclair, 2005). →

Requisitos para a elaboração de um corpus computadorizado

Daí a importância de se fazerem escolhas adequadas, de modo que o corpus possa de fato espelhar comportamentos linguísticos. Questões que devem ser feitas durante a seleção dos textos são: quais documentos? Quais tipos de textos? Quais gêneros textuais? Enfim, o que de fato representa os usos linguísticos de uma comunidade?

A representatividade

A característica mais facilmente associada à representatividade é justamente a extensão do corpus, o que significa em termos simples que para ter representatividade o corpus deve ser o maior possível (Sinclair, 1991, *apud* BERBER SARDINHA, 2000).

Tamanho de corpus

Segundo a **abordagem histórica**, proposta por Berber Sardinha (2003), a classificação geral referente ao tamanho de corpus é a seguinte:

Tamanho em palavras	Classificação
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Compilação, gerenciamento e manipulação

Seleção dos textos

1. estudo exploratório dos textos existentes bem como dos gêneros aos quais eles pertencem.

Seleção dos textos

1. estudo exploratório dos textos existentes bem como dos gêneros aos quais eles pertencem.

Seleção dos textos

1. estudo exploratório dos textos existentes bem como dos gêneros aos quais eles pertencem.

Seleção dos textos

1. estudo exploratório dos textos existentes bem como dos gêneros aos quais eles pertencem.

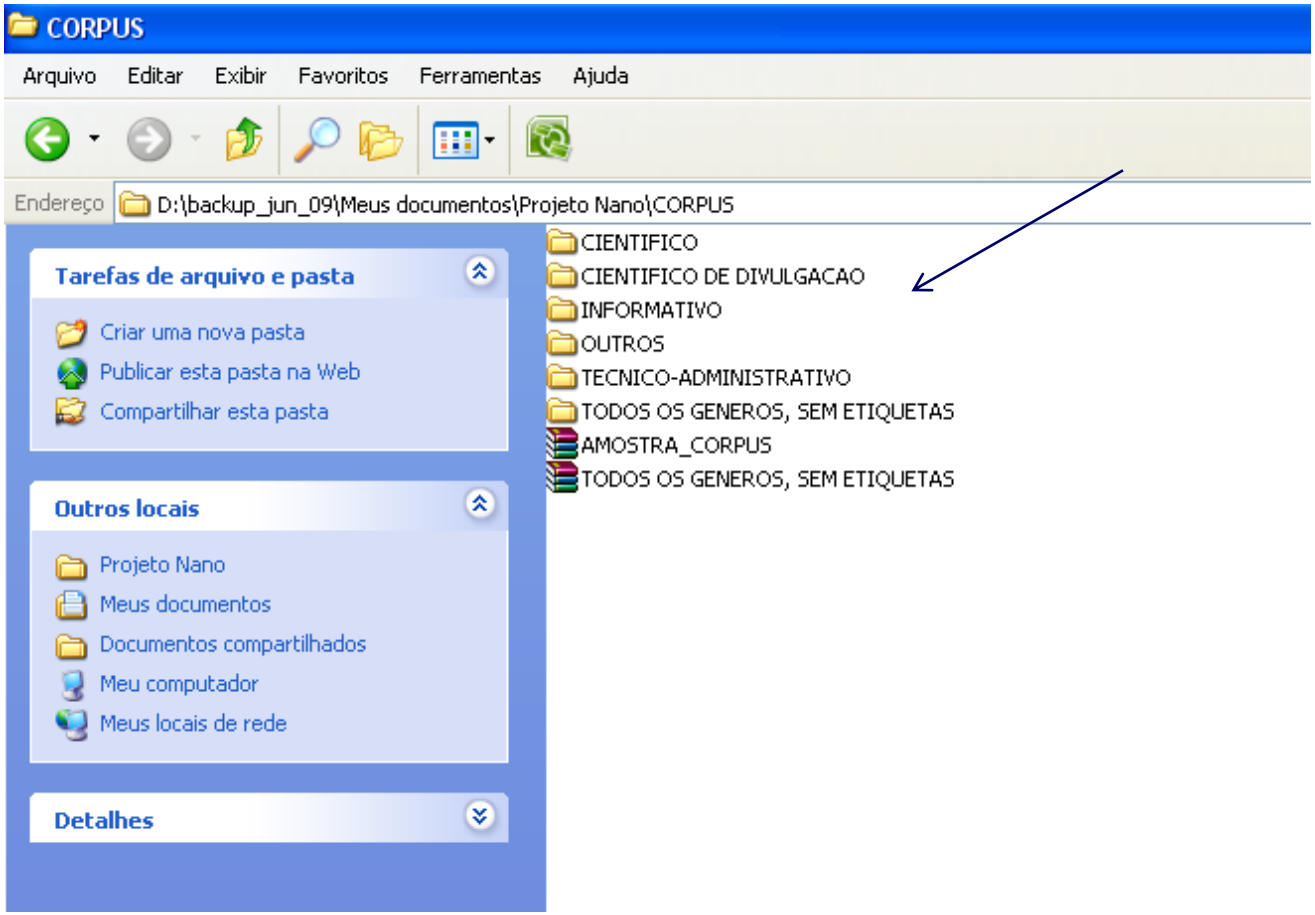
Seleção dos textos

1. estudo exploratório dos textos existentes bem como dos gêneros aos quais eles pertencem.

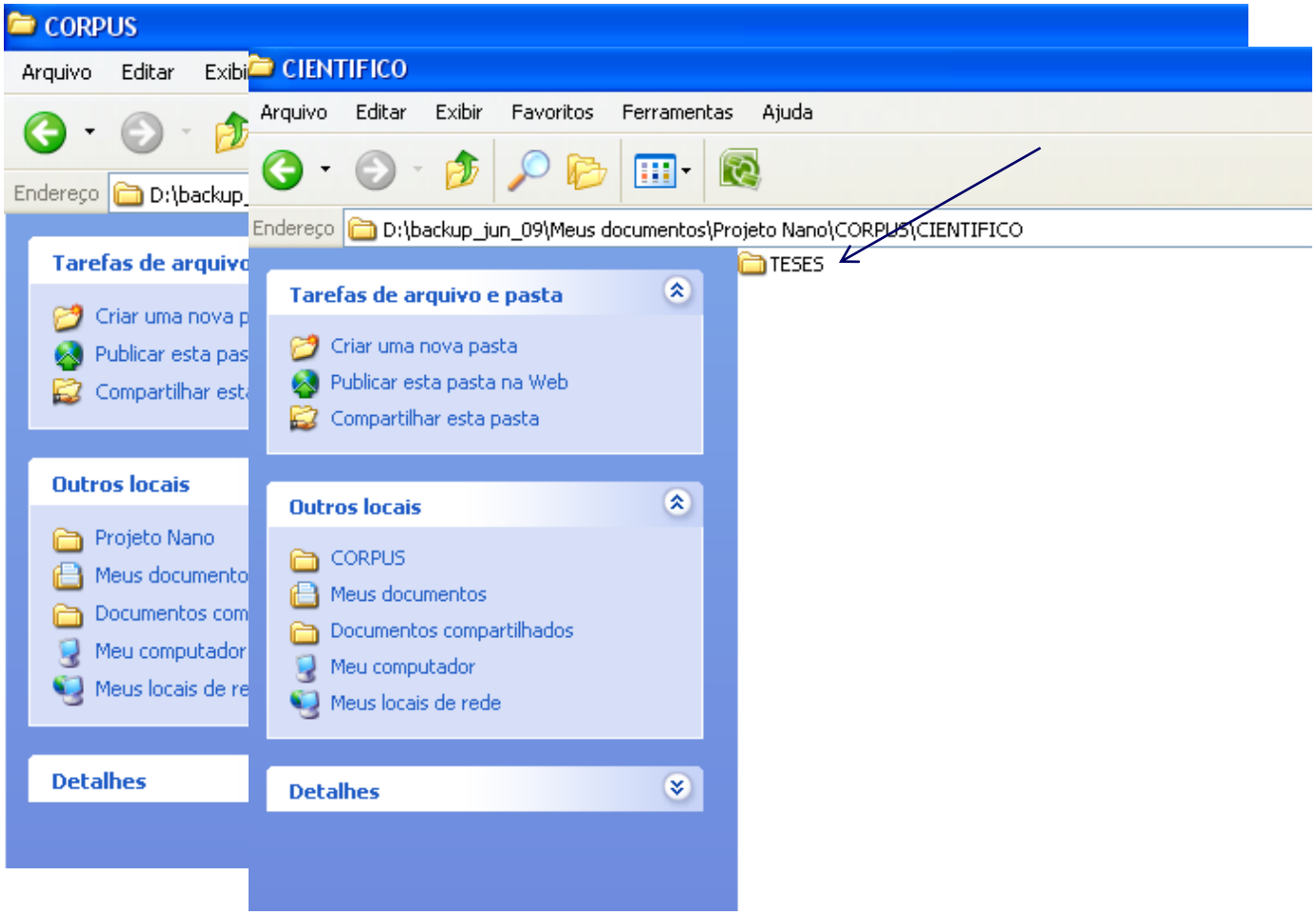
Compilação

2. Consiste no **armazenamento em arquivos predeterminados** de todos os textos pertinentes e relevantes para a pesquisa.

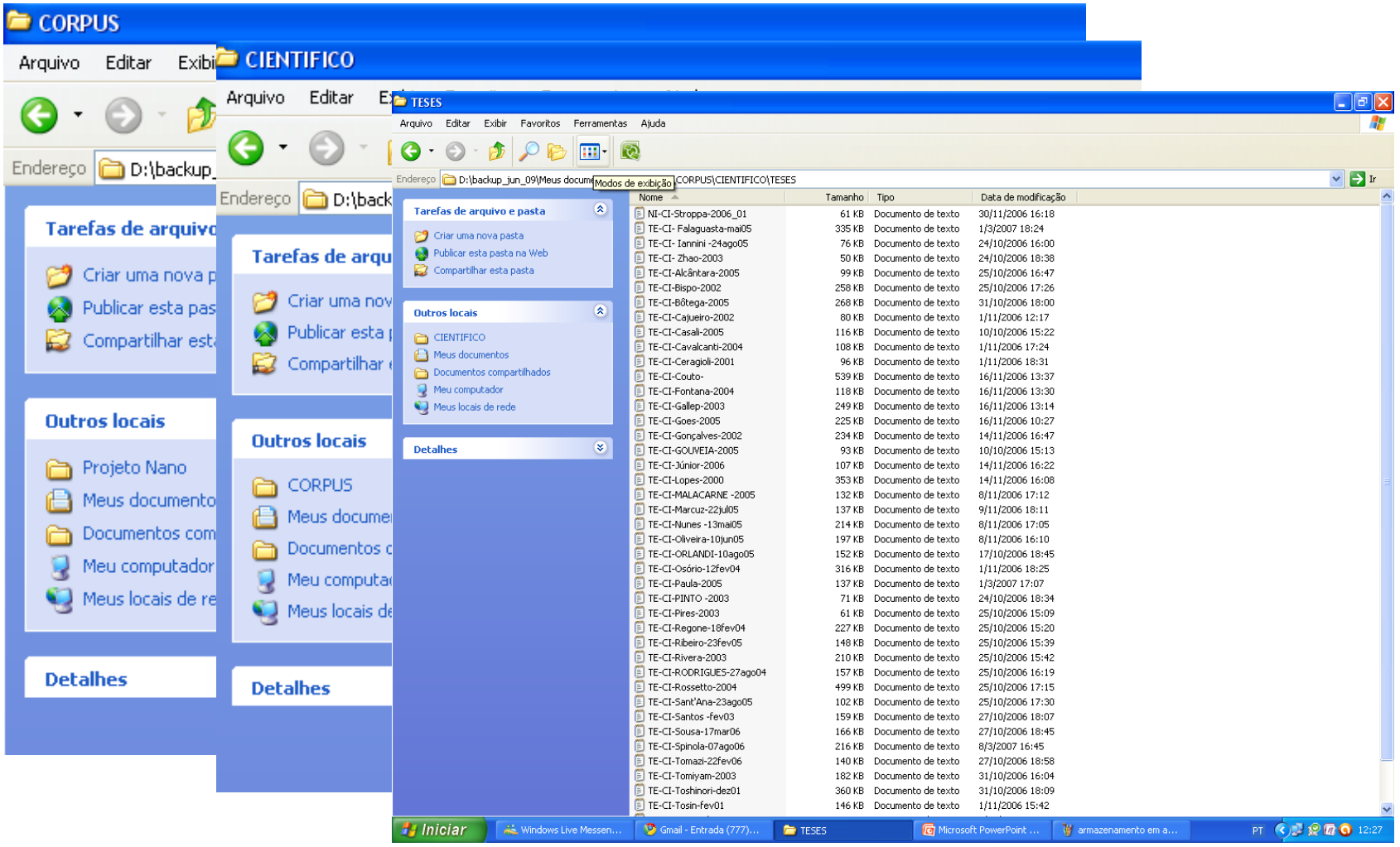
Armazenamento em arquivos



Armazenamento em arquivos



Armazenamento em arquivos



Manipulação do corpus

2. Constitui na **conversão, limpeza** e **formatação**, de maneira a preparar o corpus para o processamento computacional.

Conversão

Praticamente todas as ferramentas computacionais operam com o formato **.txt** (=bloco de notas)

Portanto,

FORMATOS ORIGINAIS	>>	FORMATO PADRÃO
Microsoft Word (“.doc”)		extensão “.txt”
HyperText Markup Language (“.html”)		
Portable Document Format (“.pdf”)		

Conversão

Praticamente todas as ferramentas computacionais operam com o formato **.txt** (=bloco de notas)

Portanto,

FORMATOS ORIGINAIS	>>	FORMATO PADRÃO
Microsoft Word (“.doc”)		extensão “.txt”
HyperText Markup Language (“.html”)		
Portable Document Format (“.pdf”)		

Conversão

Praticamente todas as ferramenta
operam com o formato .txt (=b

Portanto,

FORMATOS ORIGINAIS >> FORMATO PADRÃO

Microsoft Word (“.doc”)

HyperText Markup Language (“.html”)

Portable Document Format (“.pdf”)

Não possui códigos de formatação específicos, apenas caracteres do teclado.

extensão “.txt”

Conversão

Para as extensões **.doc** e **.html** o procedimento usual é o "copia-e-cola".

Para **.pdf**: a conversão pode ser automática:

utilizando-se o programa "XPDF"

↓

http://www.foolabs.com/xpdf/index.html

Conversão

O programa XPDF está disponível apenas em ambiente Linux, ou em ambientes Linux emulado em Windows.

Para isso, pode-se utilizar o

CYGWIM >> EMULADOR DE AMBIENTE LINUX

http://www.cygwin.com/setup.exe

Conversão

O programa XPDF está disponível apenas em ambiente Linux, ou em ambientes Linux emulado em Windows.

Para isso, pode-se utilizar o

CYGWIM >> EMULADOR DE AMBIENTE LINUX

http://www.cygwin.com/setup.exe

Limpeza e formatação

- **Limpeza**: excluir tabelas, gráficos, fórmulas, cálculos, imagens, números de página, referências bibliográficas, enfim, toda a informação que não esteja sob a forma de texto.

- **Formatação:** formatar cada texto no modo desejado para a pesquisa.

Nomeação de arquivos e geração de cabeçalhos

Depois dos textos convertidos em formato **.txt**, limpos e formatados, eles devem receber um nome.

Essa nomeação deve seguir determinado padrão, de forma a facilitar a recuperação posterior de cada texto.

Nomeação de arquivos e geração de cabeçalhos

Após a nomeação dos arquivos, é gerado (de forma semiautomática) um cabeçalho para cada texto (versão adaptada do *Editor de cabeçalho* do Projeto Lacio-Web

(http://www.nilc.icmc.usp.br/lacioweb/)

Editor de cabeçalho (1)

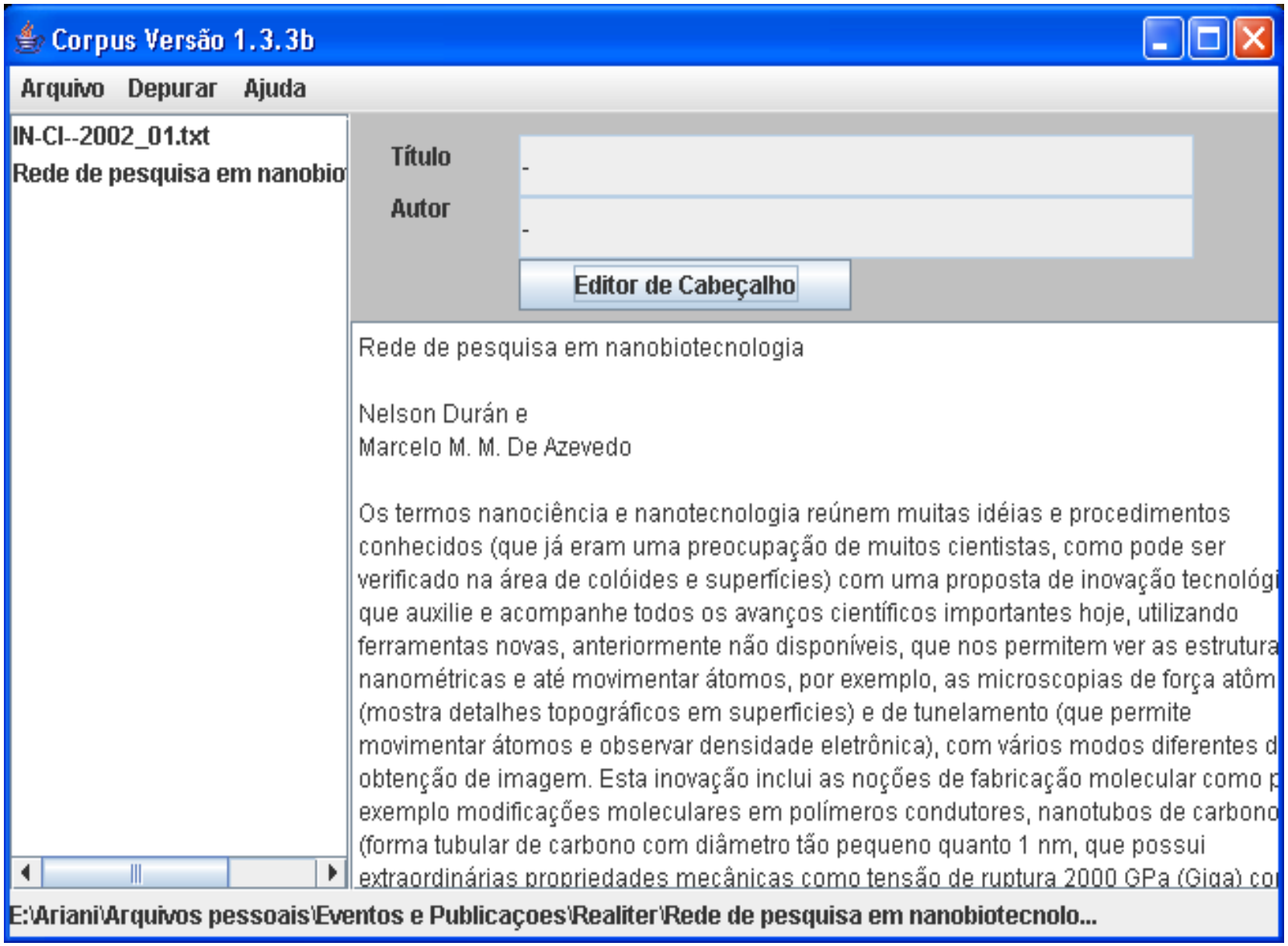


Figura 1: Editor de cabeçalho (adaptado) do Projeto Lacio-Web.

Editor de cabeçalho (2)

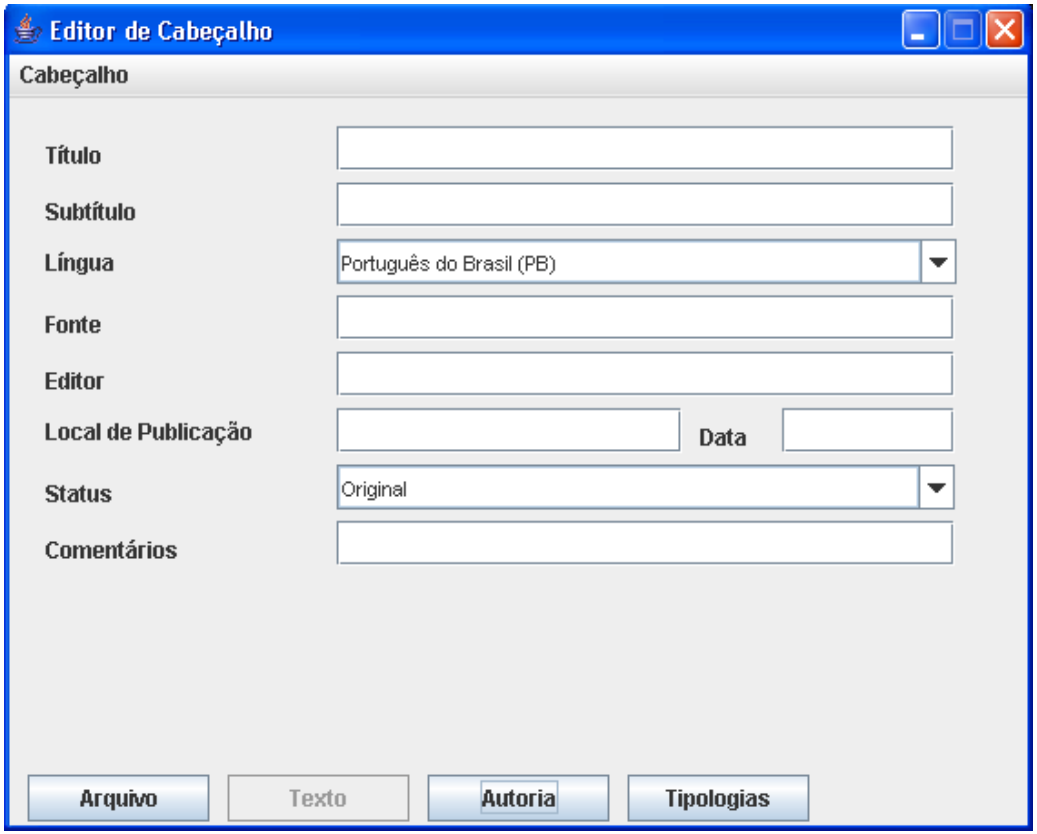


Figura 2: Pop-ups do Editor para a especificação de diversas informações.

Editor de cabeçalho (2)

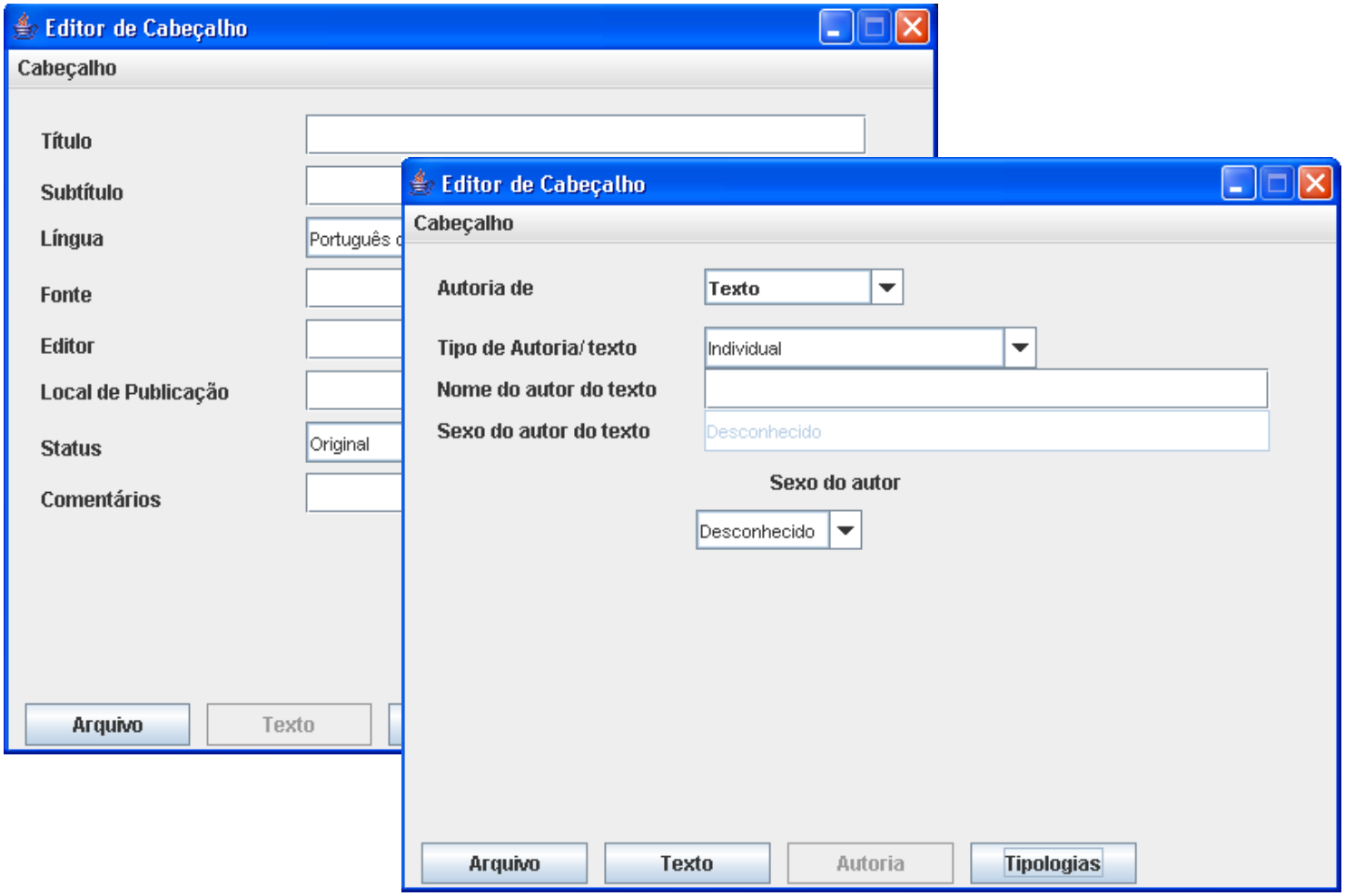


Figura 2: Pop-ups do Editor para a especificação de diversas informações.

Editor de cabeçalho (2)

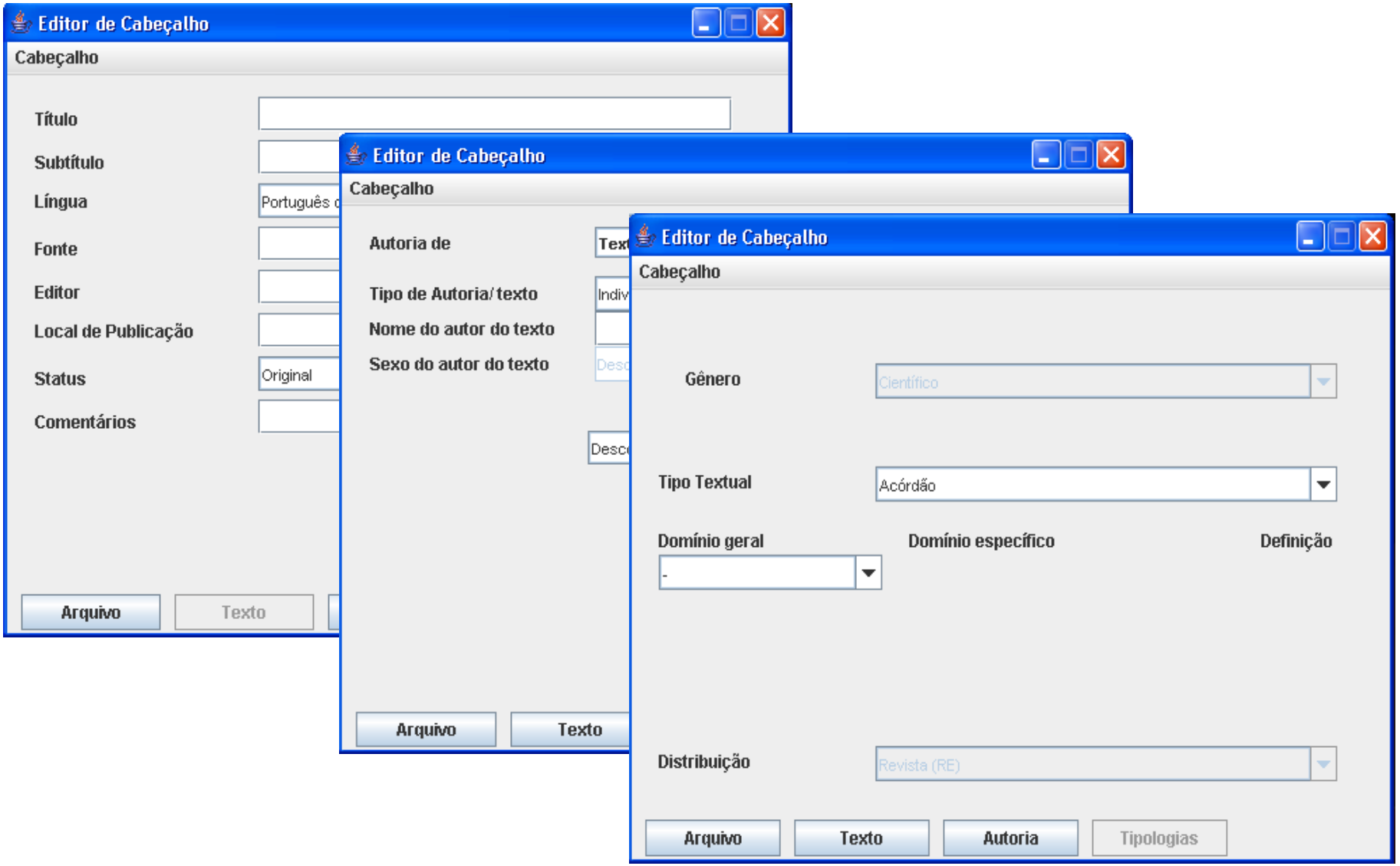


Figura 2: Pop-ups do Editor para a especificação de diversas informações.

Editor de cabeçalho (3)

IN-CI--2002_01.txt - Bloco de notas

Arquivo Editar Formatar Exibir Ajuda

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- v:1.3.3b-->
<text>
<header>
        <title>
                <fileName>IN-CI--2002_01.txt</fileName>
                <corpus>Referência</corpus>
                <nPages>6</nPages>
                <nWords>1939</nWords>
                <sample>Integra</sample>
        </title>
        <sourceText>
                <titleOfText>Rede de pesquisa em nanobiotecnologia</titleOfText>
                <language>Português do Brasil (PB)</language>
                <source>Revista Com Ciência</source>
                <editor>Revista Com Ciência</editor>
                <pubPlace>Brasil</pubPlace>
                <date>2002</date>
                <status>Original</status>
                <comments>Caderno "Caderno Nanociência e Nanotecnologia".</comments>
                <accessDate>25.05.06</accessDate>
                <site>http://www.comciencia.br/reportagens/nanotecnologia/nano20.htm</site>
        </sourceText>
<author>
        <textAuthor>
                <name>Nelson Durán e  Marcelo M. M. De Azevedo</name>
                <gender>Masculino</gender>
                <type>Múltiplo</type>
        </textAuthor>
</author>
        <textClassification>
                <textGenre>
                        <genre>Científico</genre>
                </textGenre>
                <textType>Artigo</textType>
                <domain>
                        <generalDomain defined="annotador-def">Generalidades</generalDomain>
                        <specificDomain>Ciência & Tecnologia</specificDomain>
                </domain>
                <distribution>Internet</distribution>
        </textClassification>
</header>
<body>
```

Figura 3: Texto (em .txt) gerado pelo Editor de cabeçalho.

Editor de cabeçalho (3)

IN-CI--2002_01.txt - Bloco de notas

Arquivo Editar Formatar Exibir Ajuda

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- v:1.3.3b-->
<text>
<header>
        <title>
                <fileName>IN-CI--2002_01.txt</fileName>
                <corpus>Referência</corpus>
                <nPages>6</nPages>
                <nWords>1939</nWords>
                <sample>Integra</sample>
        </title>
        <sourceText>
                <titleOfText>Rede de pesquisa em nanobiotecnologia</titleOfText>
                <language>Português do Brasil (PB)</language>
                <source>Revista Com Ciência</source>
                <editor>Revista Com Ciência</editor>
                <pubPlace>Brasil</pubPlace>
                <date>2002</date>
                <status>Original</status>
                <comments>Caderno "Caderno Nanociência e Nanotecnologia".</comments>
                <accessDate>25.05.06</accessDate>
                <site>http://www.comciencia.br/reportagens/nanotecnologia/nano20.htm</site>
        </sourceText>
<author>
        <textAuthor>
                <name>Nelson Durán e  Marcelo M. M. De Azevedo</name>
                <gender>Masculino</gender>
                <type>Múltiplo</type>
        </textAuthor>
</author>
        <textClassification>
                <textGenre>
                        <genre>Cientifico</genre>
                </textGenre>
                <textType>Artigo</textType>
                <domain>
                        <generalDomain defined="annotador-def">Generalidades</generalDomain>
                        <specificDomain>Ciência & Tecnologia</specificDomain>
                </domain>
                <distribution>Internet</distribution>
        </textClassification>
</header>
<body>
```

etiquetas XML

Figura 3: Texto (em .txt) gerado pelo Editor de cabeçalho.

Philologic: o que é?

Conjunto de softwares desenvolvido pelo Projeto ARTFL e pelo Digital Library Development Center (DLDC) da Universidade de Chicago.

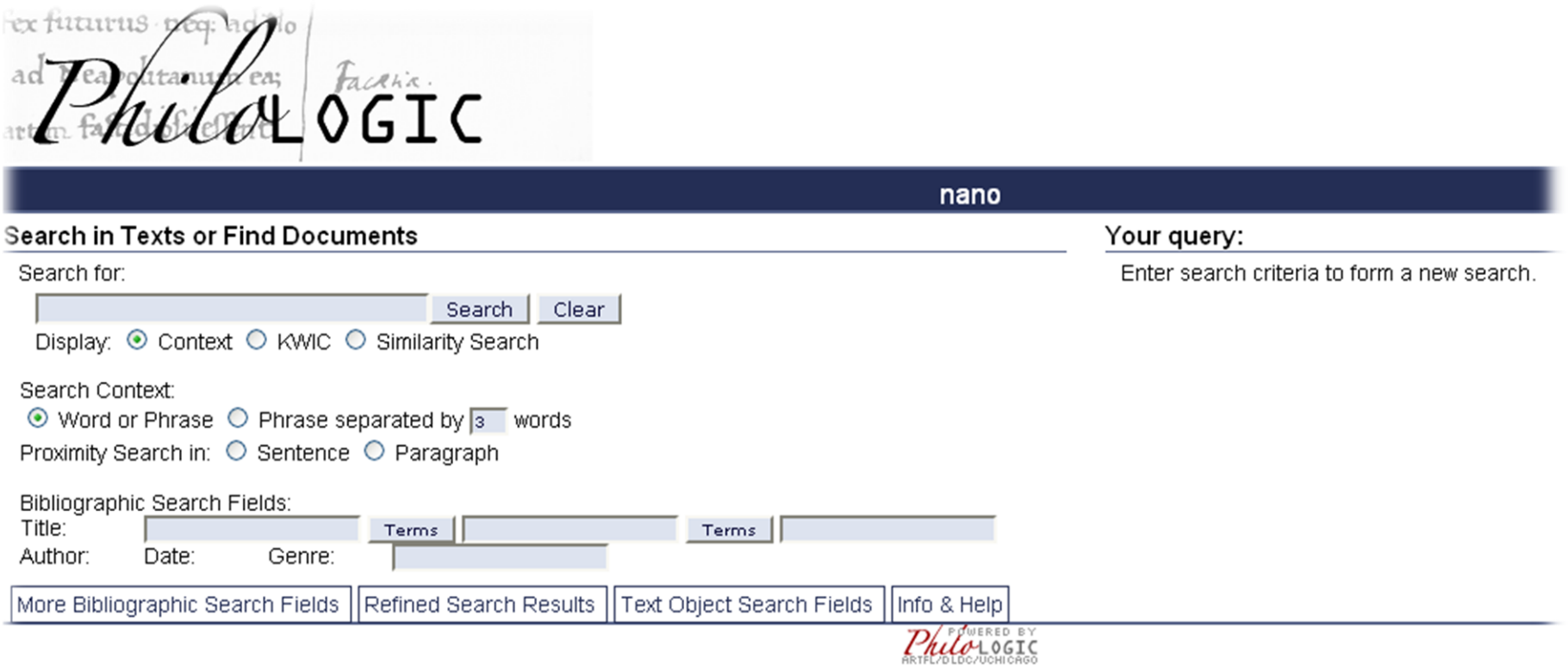
Foi elaborado para lidar com grandes quantidades de documentos codificados (em XML ou SGML), permitindo a realização de buscas sofisticadas, buscas de textos completos e recuperação de metadados.

Philologic: vantagens

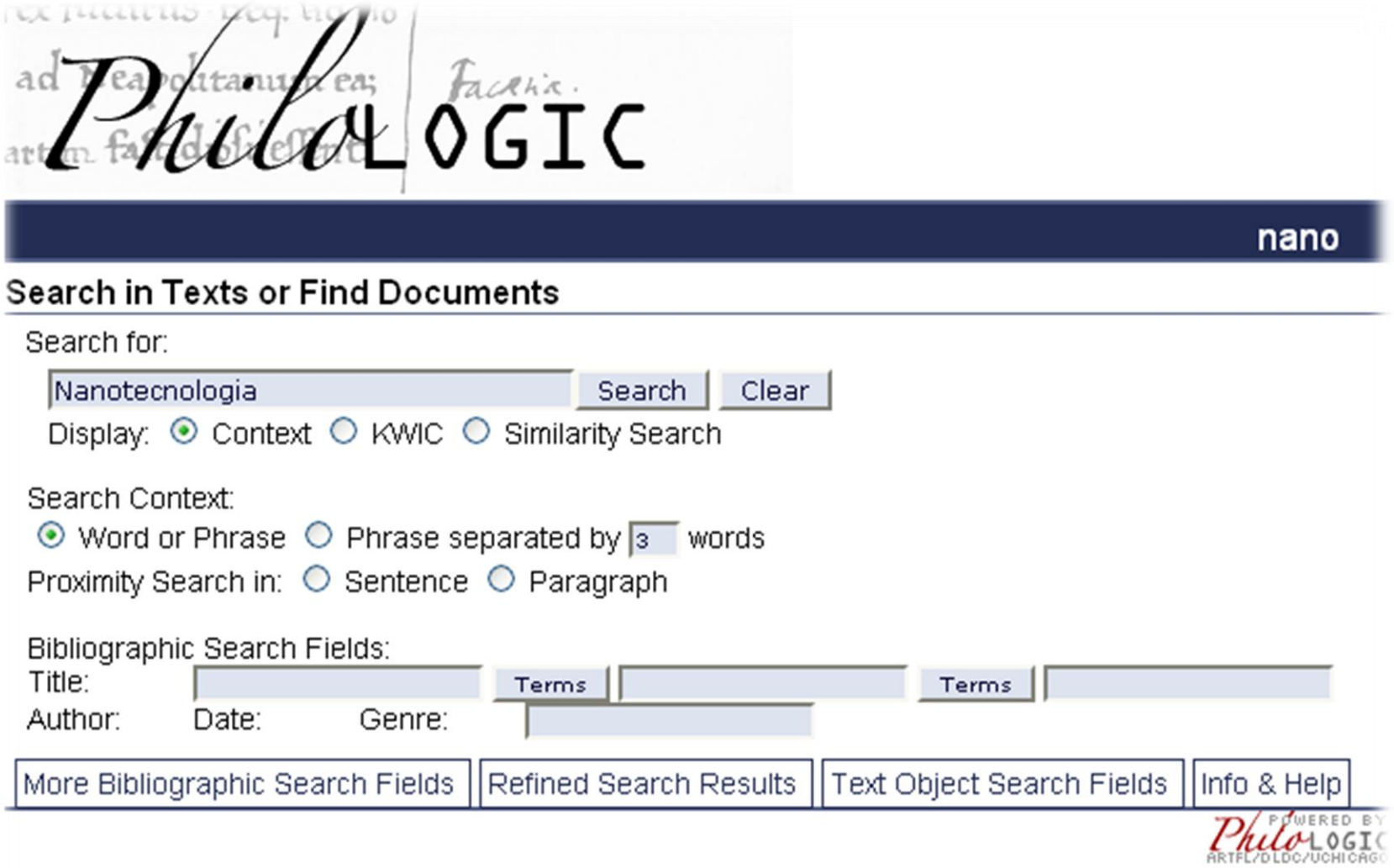
Software de código livre compatível com outros recursos computacionais usado por numerosas instituições acadêmicas e por organizações comerciais.

Implementação on-line, possibilitando que diferentes usuários de uma mesma pesquisa possam manipular o corpus a partir de diferentes locais sem dificuldades.

http://www.lib.uchicago.edu/efts/ARTFL/philologic/



http://www.lib.uchicago.edu/efts/ARTFL/philologic/



Em destaque o campo “Search For” e opções de busca

http://www.lib.uchicago.edu/efts/ARTFL/philologic/

3. KÍRIAN PIMENTA... . *DESENVOLVIMENTO DE PIGMENTO*... [Paragraph | Section]

e centros dedicados à nanociência. Várias reuniões têm sido efetuadas entre representantes ambientais, industriais e políticos de forma a ponderar futuras medidas relativas à nanotecnologia. Enquanto isso, a comunidade científica prepara-se para argumentar sobre os prós e contras da **nanotecnologia**. Encontram-se no mercado alguns produtos nanotecnológicos, como exemplo, tem-se a comercialização de nanopartículas de óxido de zinco para produção de cremes protetores solares, as quais não refletem a luz solar e tornam o creme transparente, em vez do branco

---

4. KÍRIAN PIMENTA... . *DESENVOLVIMENTO DE PIGMENTO*... [Paragraph | Section]

O controle da matéria em nanoescala apresenta um importante papel em diversas áreas científicas, como física, química, ciência dos materiais, biologia, medicina, engenharia e computacão. Portanto, uma vez determinadas as novas propriedades dos materiais e sistemas nesta escala, a **nanotecnologia** pode causar impacto na produção de quaisquer objetos, desde automóveis, pneus, até circuitos de computadores e utensílios da medicina10. 1.3 Estrutura do Titanato de Níquel Óxidos que apresentam fórmula geral ABO3 tem sido considerados como pertencentes ao

---

5. Marcos Pivetta. *Arquitetos de moléculas* [Paragraph | Section]

Arquitetos de moléculas Pesquisa que intervém na estrutura de blocos ínfimos de matéria produz compostos como um sensor para conservante de vinhos Marcos Pivetta No século 21, o mundo da ciência vai ficar menor, segundo os que se dedicam à emergente área da **nanotecnologia** molecular. Os especialistas desse ramo das nanociências propõem-se a dominar a manipulação das moléculas e da menor partícula de matéria capaz de conservar as características químicas de um elemento- o átomo. Esse é o propósito de Henrique Toma, do Instituto de Química

---

6. Marcos Pivetta. *Arquitetos de moléculas* [Paragraph | Section]

nelas alguma reação química presente na natureza ou no corpo humano, como a fotossíntese- em que a planta usa a luz para converter água, dióxido de carbono e minerais em oxigênio e em compostos ricos em energia- ou as decorrentes da ação de enzimas. Em tese, o controle pleno da **nanotecnologia** molecular, um sonho ainda longe de ser alcançado, permitiria ao homem rearranjar blocos ínfimos de matéria como bem entendesse. E, assim, refazer moléculas existentes ou criar novas. "Quase não há campo da atividade humana em que a nanotecnologia molecular não possa ser útil

---

7. Marcos Pivetta. *Arquitetos de moléculas* [Paragraph | Section]

de enzimas. Em tese, o controle pleno da nanotecnologia molecular, um sonho ainda longe de ser alcançado, permitiria ao homem rearranjar blocos ínfimos de matéria como bem entendesse. E, assim, refazer moléculas existentes ou criar novas. "Quase não há campo da atividade humana em que a **nanotecnologia** molecular não possa ser útil ao homem, desde a produção de alimentos até o tratamento de doenças", diz Toma. Vinho equilibrado Um dos compostos que ele mais usa são as porfirinas, tipo de pigmento abundante na natureza e que atua em vários processos

---

8. Marcos Pivetta. *Arquitetos de moléculas* [Paragraph | Section]

usa nanopartículas de carbono para reforçar a borracha de seu produto. "O mundo nano está aí. A gente é que ainda não se deu conta disso", diz Elson Longo, pesquisador do Departamento de Química da Universidade Federal de São Carlos (UFSCar). As bases da noção moderna de **nanotecnologia** molecular, no entanto, são mais recentes. No fim de 1959, na reunião anual da Sociedade Americana de Física, Richard P. Feynman fez um discurso provocadorque entraria para a história como o pontapé inicial. "Por que não podemos escrever todos os 24 volumes da Enciclopédia

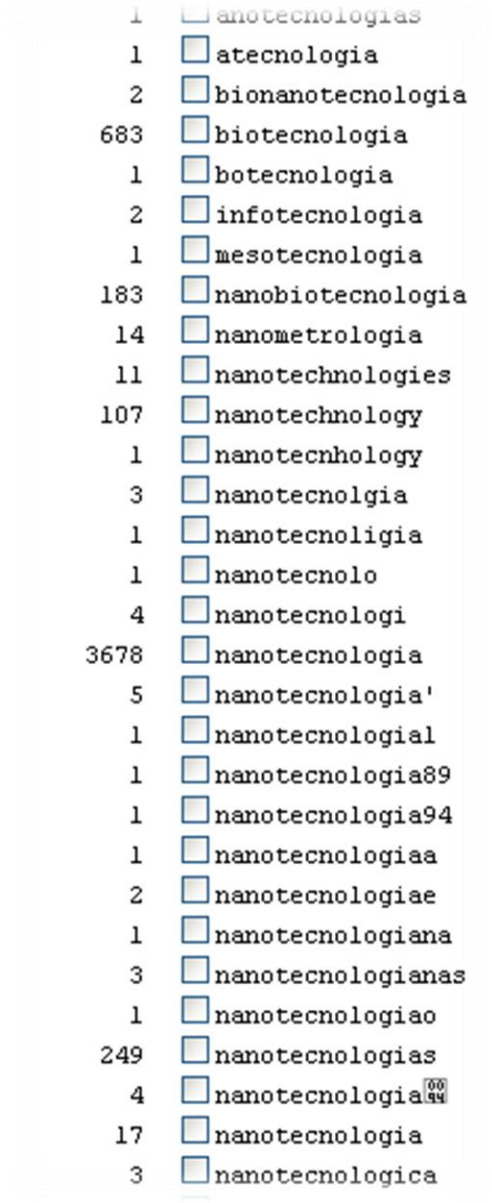
Exibição de resultado no formato contexto expandido

http://www.lib.uchicago.edu/efts/ARTFL/philologic/

2. **0990** (bib:p.0)ponderar futuras medidas relativas à **nanotecnologia**. Enquanto isso, a comunidade cient
3. **0990** (bib:p.0)rgumentar sobre os prós e contras da **nanotecnologia**. Encontram-se no mercado alguns pr
4. **0990** (bib:p.0)materiais e sistemas nesta escala, a **nanotecnologia** pode causar impacto na produção
5. **0335** (bib:p.0)s que se dedicam à emergente área da **nanotecnologia** molecular. Os especialistas desse
6. **0335** (bib:p.0)nzimas. Em tese, o controle pleno da **nanotecnologia** molecular, um sonho ainda longe de
7. **0335** (bib:p.0)á campo da atividade humana em que a **nanotecnologia** molecular não possa ser útil ao
8. **0335** (bib:p.0)FSCar). As bases da noção moderna de **nanotecnologia** molecular, no entanto, são mais r
9. **0335** (bib:p.0) de forma precária. Se os adeptos da **nanotecnologia** molecular fossem astrônomos, segu
10. **0336** (bib:p.0)ara lançar seu Programa Nacional de **Nanotecnologia**. Em 2001, só o Japão pretende ga
11. **0336** (bib:p.0)de um possível programa nacional de **nanotecnologia**. Compareceram 32 pesquisadores. No
12. **0336** (bib:p.0)es da ciência nacional. Sabemos que **nanotecnologia** é um setor emergente e muito impo
13. **0357** (bib:p.0)cnologia Internet O grande mundo da **nanotecnologia** Daqui a poucos anos estaremos vive
14. **1015** (bib:p.0)aro. Experiências para a introdução **nanotecnologia** [16] 3 Ainda que a origem da lumi
15. **0340** (bib:p.0)a investir em pesquisas na áreas de **nanotecnologia**, computadores mais potentes e simu
16. **0121** (bib:p.0)s de medicação para uso no reino da **nanotecnologia**."A beleza deste sistema é que, cr
17. **0345** (bib:p.0)decidiram investir em nanociência e **nanotecnologia**- a habilidade de trabalhar átomo
18. **0145** (bib:p.0)dades Ciência & Tecnologia Internet **Nanotecnologia**: só o orçamento é grande Da red
19. **0145** (bib:p.0)aboratórios, em busca do domínio da **nanotecnologia**. George Bush, na ânsia do corte d
20. **0145** (bib:p.0) 30 grupos de pesquisas cuidando de **nanotecnologia** nos Estados Unidos. E não parece
21. **0145** (bib:p.0)), o total de recursos aplicados em **nanotecnologia** no mundo todo mais do que dobrou n
22. **0341** (bib:p.0)ogia da informação, meio ambiente e **nanotecnologia**-, além da reforma do sistema de C
23. **0475** (bib:p.0)& Tecnologia Internet Nanociência e **Nanotecnologia** Alaor Chaves Há muito sabemos que
24. **0475** (bib:p.0)temente, surgiram a nanociência e a **nanotecnologia** (N & N), que têm por meta dominar
25. **0475** (bib:p.0)anômetros. Assim, a nanociência e a **nanotecnologia** visam, respectivamente, a compreen

Resultados de busca exibidos no formato "KWIC"

http://www.lib.uchicago.edu/efts/ARTFL/philologic/



Detecção automática de termos semelhantes (recurso *Similarity Search*)

Unitex: o que é?

Desenvolvido na Universidade Marne-La-Vallée (França) por Sébastien Paumier (PAUMIER, 2002), o Unitex consiste num conjunto de programas que permite o processamento de grandes quantidades de textos, em diversas línguas.

Unitex: o que é?

Na versão 2.0, o Unitex tem módulos para o alemão, coreano, espanhol, finlandês, francês, grego antigo, grego moderno, inglês, italiano, norueguês, polonês, português do Brasil, português europeu, russo, sérvio (tanto com o alfabeto cirílico quanto com o latino) e tailandês.

Unitex: vantagens

O Unitex funciona com base em dicionários eletrônicos de cada uma das línguas que o integram.

Para o português do Brasil, o Unitex traz um dicionário eletrônico bastante extenso:

- cerca de 67.500 formas canônicas (ou lemas)
- 880 mil formas flexionadas
- 4.500 formas compostas com hífen

Unitex: vantagens

O Unitex funciona com base em dicionários eletrônicos de cada uma das línguas que o integram.

Para o português do Brasil, o Unitex traz um dicionário eletrônico bastante extenso:

- cerca de 67.500 formas canônicas (ou lemas)
- 880 mil formas flexionadas
- 4.500 formas compostas com hífen

Unitex: vantagens

Elaborado por Muniz (2004) ,
a partir do léxico do Núcleo
Interinstitucional de Linguística
Computacional (NILC), sediado na
Universidade de São Paulo (USP),
campus de São Carlos (SP, Brasil).

dicionários eletrônicos
ntegram.

ex traz um **dicionário**

s (ou lemas)

4.500 formas compostas com hífen

Unitex: vantagens

O programa também permite que qualquer usuário crie seus próprios dicionários, integrando novas unidades lexicais ou, ainda, acrescentando novas informações morfológicas, sintáticas e semânticas ao léxico já existente ou ainda gerando novas formas a partir de uma forma canônica.

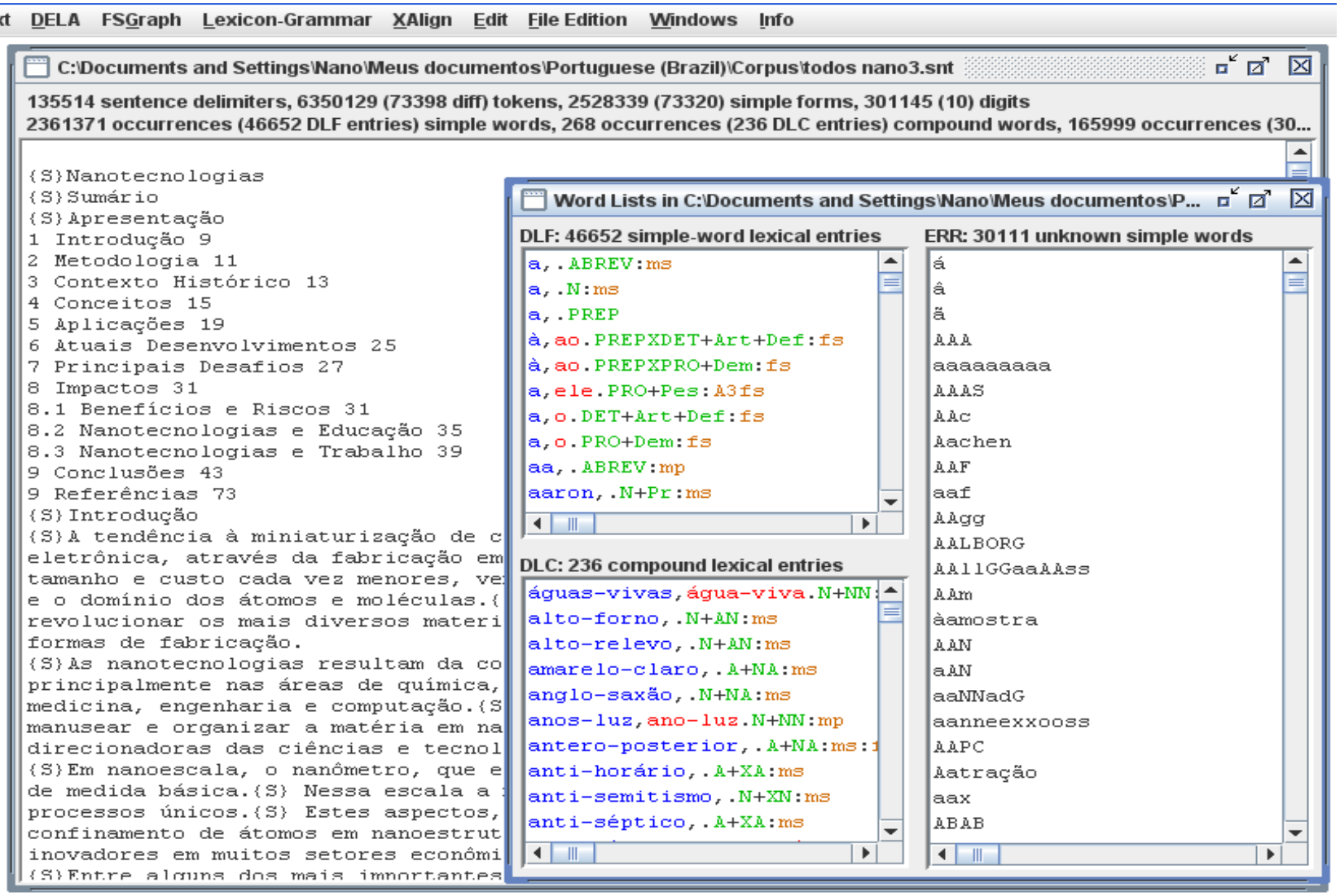
Unitex: vantagens

Esses dicionários possibilitam ao usuário do programa a realização de buscas pela forma exata, pela forma canônica e também pelas categorias gramaticais. Além disso, o programa permite a combinação desse tipo de busca com a busca por formantes. Essas características fazem com que o Unitex possa ser particularmente útil em buscas de construções complexas.

Unitex: vantagens

Outra característica dessas buscas é o fato de elas poderem ser realizadas tanto por expressões regulares quanto por grafos, os quais podem ser desenhados pelos utilizadores

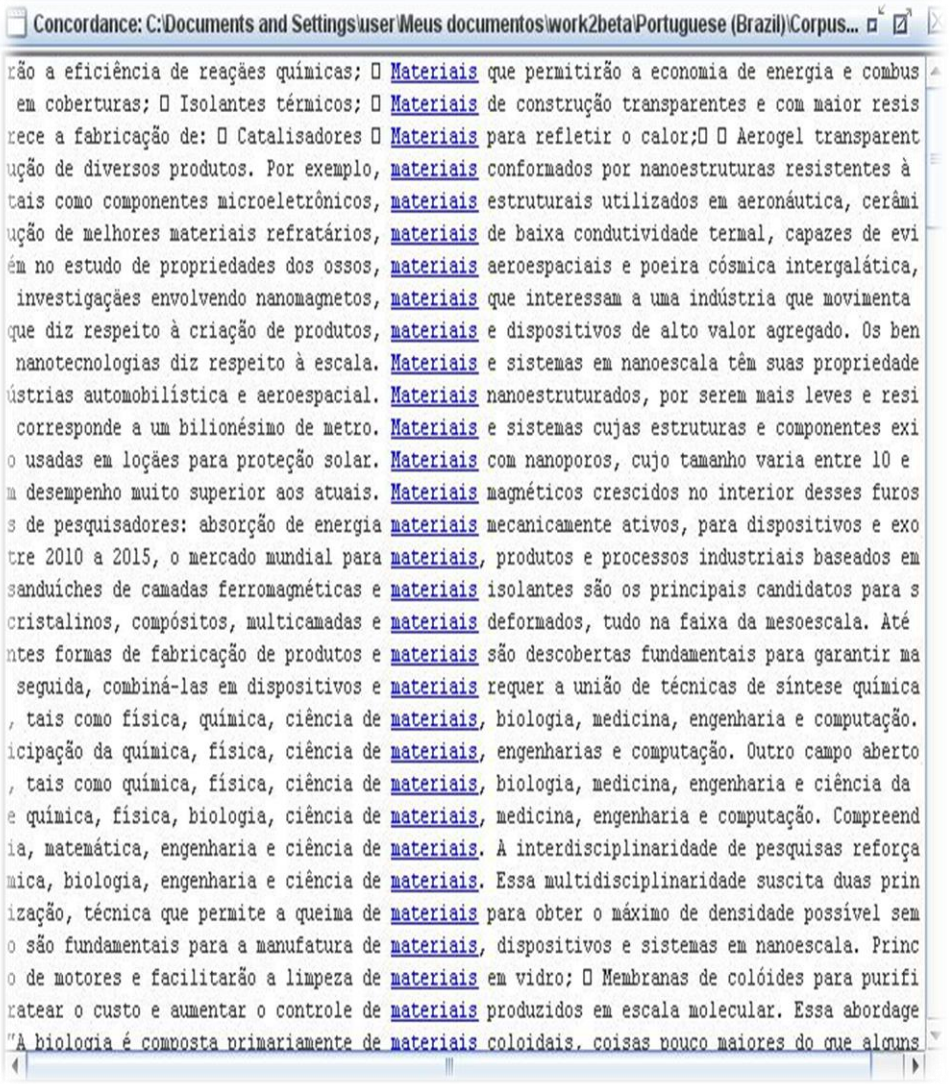
http://www-igm.univ-mlv.fr/~unitex/



Corpus depois de processado

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

Resultado das concordâncias com o item léxico *material*



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

Possibilidades de busca:

<material><A>: **1.989 ocorrências**
<material><!DIC>: **722 ocorrências**
<material>de<!DIC><A>: **2 ocorrências**
<material>de<N>: **353 ocorrências**
<material>de<N><A>: **127 ocorrências**
<material>de<N><!DIC>: **3 ocorrências**

(<material>).(<A>+<!DIC>+de<!DIC>+de<N>+de<N><!DIC>+de<!DIC><A>+de<N><A>)

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

Possibilidades de busca:

Tamanho do corpus:

2.565.490 palavras

<material><A>: **1.989 ocorrências**

<material><!DIC>: **722 ocorrências**

<material>de<!DIC><A>: **2 ocorrências**

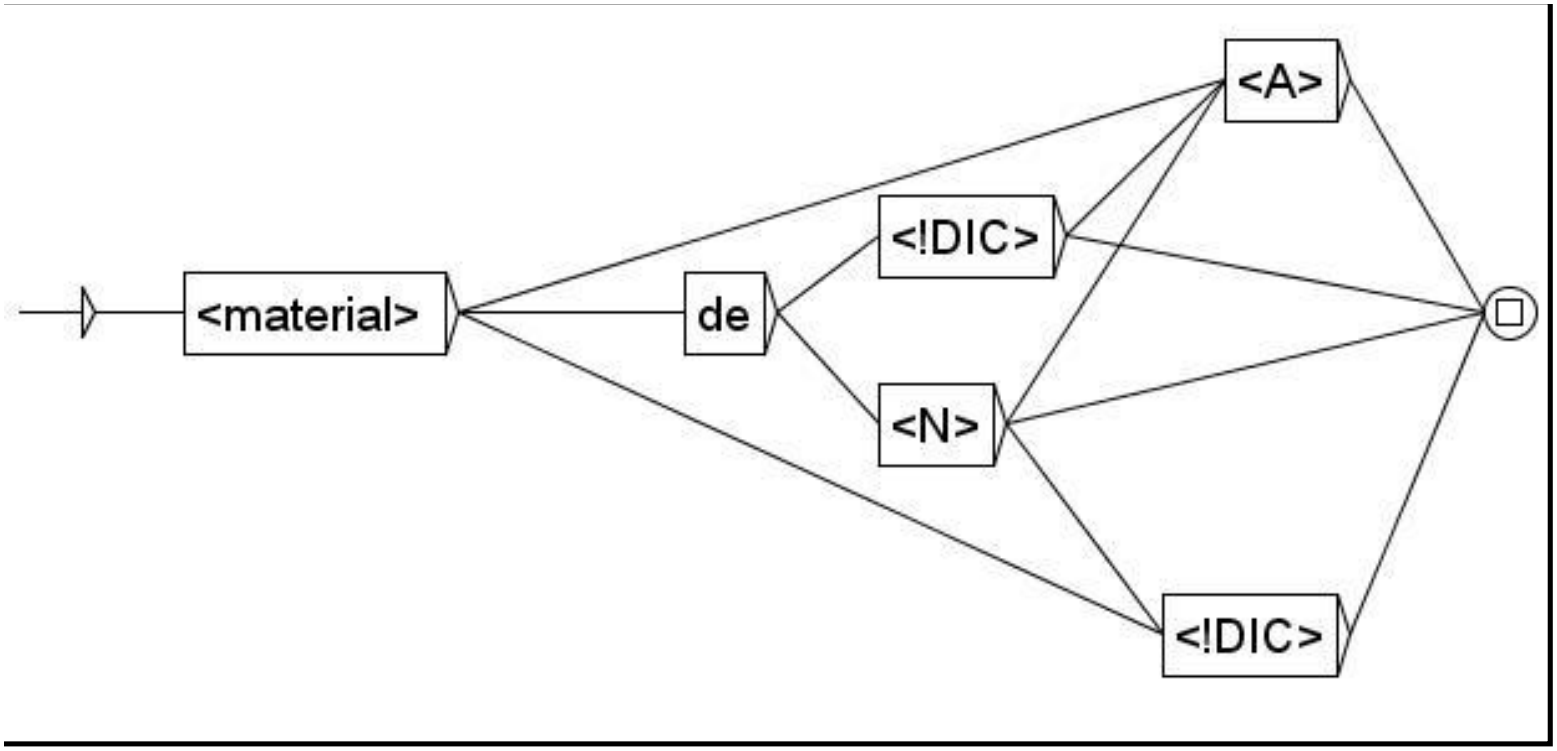
<material>de<N>: **353 ocorrências**

<material>de<N><A>: **127 ocorrências**

<material>de<N><!DIC>: **3 ocorrências**

(<material>).(<A>+<!DIC>+de<!DIC>+de<N>+de<N><!DIC>+de<!DIC><A>+de<N><A>)

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



Grafo de busca das combinações com *material*

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

Sintaxes de busca	Ocorrências
(<**processo**>).(<A>+<!DIC>+de<!DIC>+de<N>+de<N><!DIC>+de<!DIC><A>+de<N><A>)	4.071
(<**sistema**>).(<A>+<!DIC>+de<!DIC>+de<N>+de<N><!DIC>+de<!DIC><A>+de<N><A>)	3.041
(<**amostra**>).(<A>+<!DIC>+de<!DIC>+de<N>+de<N><!DIC>+de<!DIC><A>+de<N><A>)	2.546
(<**estrutura**>).(<A>+<!DIC>+de<!DIC>+de<N>+de<N><!DIC>+de<!DIC><A>+de<N><A>)	2.099
(<**propriedade**>).(<A>+<!DIC>+de<!DIC>+de<N>+de<N><!DIC>+de<!DIC><A>+de<N><A>)	2.012
(<**tecnologia**>).(<A>+<!DIC>+de<!DIC>+de<N>+de<N><!DIC>+de<!DIC><A>+de<N><A>)	1.051

WordSmith Tools: o que é?

De autoria de Mike Scott, da Aston University (Birmingham), o programa é composto basicamente de ferramentas (*Wordlist*, *Keywords* e *Concord*), e em cada uma dessas ferramentas há uma gama de outros recursos que auxiliam na descrição linguística.

WordSmith Tools: vantagens

- bom desempenho estatístico: por meio da ferramenta *Wordlist* é possível recuperar informações acerca da quantidade de palavras, sentenças e índice de riqueza vocabular do corpus;

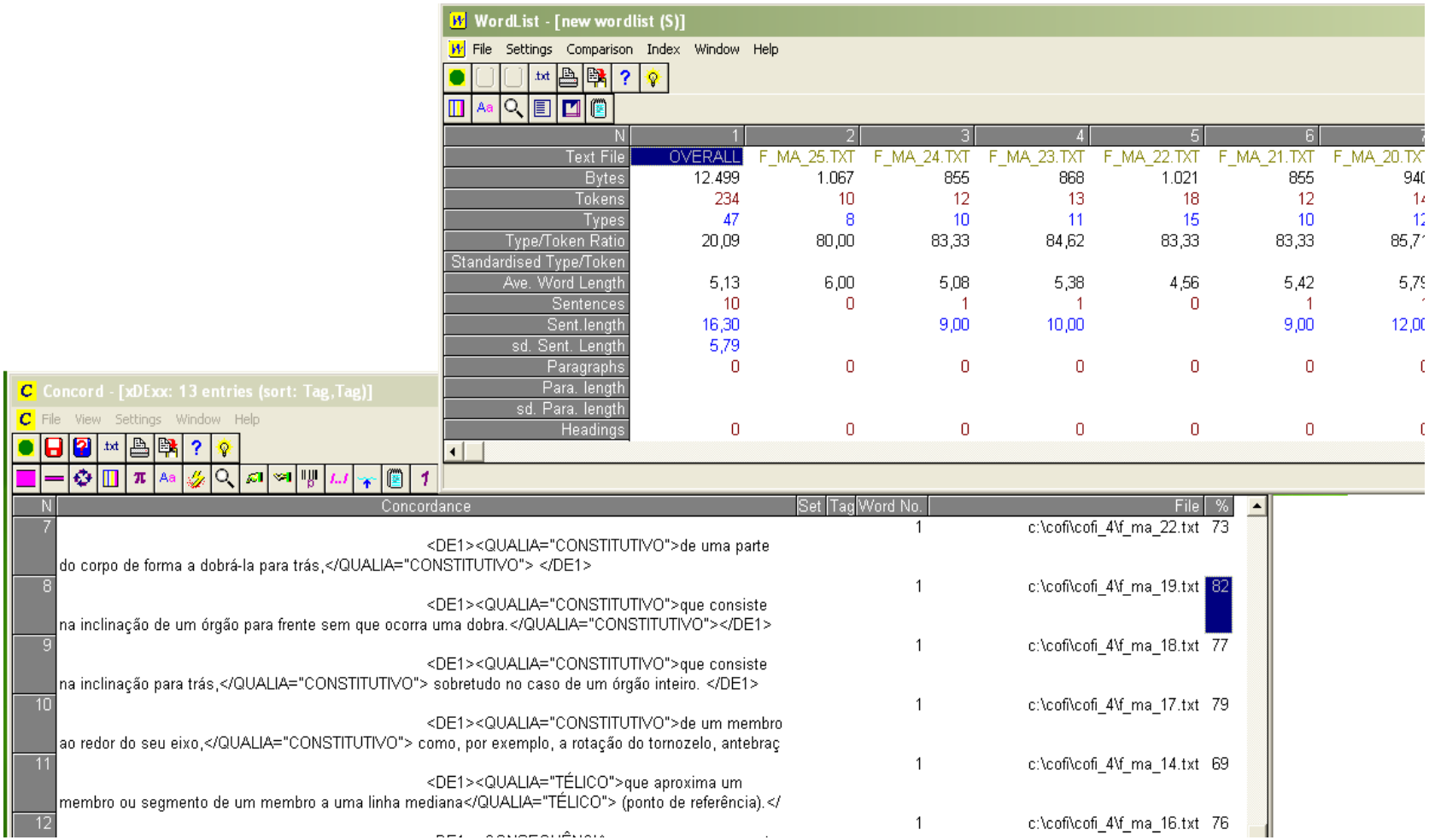
WordSmith Tools: vantagens

- o programa permite manipular vários arquivos simultaneamente, fornecendo informações (estatísticas e linguísticas) sobre cada arquivo em específico;

WordSmith Tools: vantagens

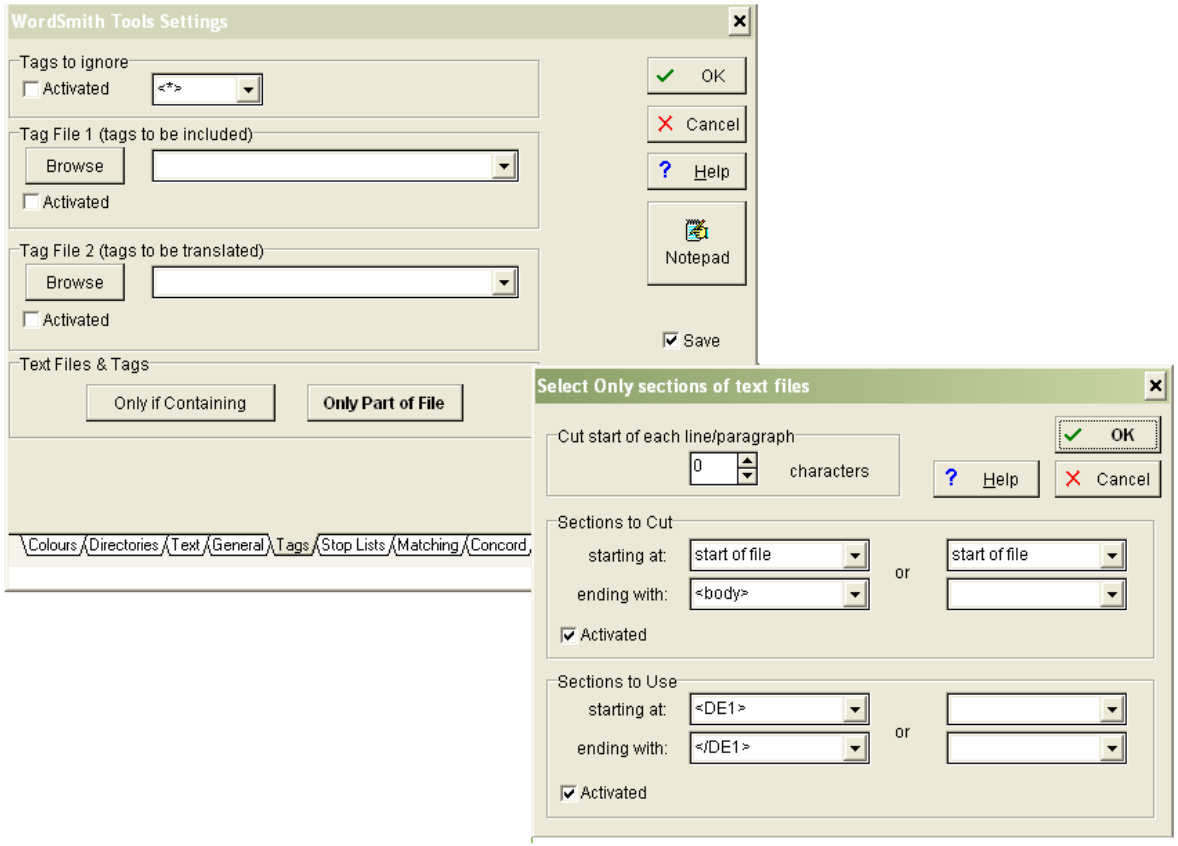
- leitura de etiquetas “xml” (*tags*): recurso que permite a leitura de etiquetas, e por conseguinte, a criação de buscas específicas.

http://www.lexically.net/wordsmith/



Telas “Concord” e “WordList”

http://www.lexically.net/wordsmith/



Recurso Tags (inserido em *WordSmith Tools Settings*), que permite fundamentalmente incluir ou não as etiquetas na visualização das ocorrências e selecionar partes do texto a serem analisadas

Corpus na pesquisa linguística:

A partir de corpus, podem-se fazer observações precisas sobre o real comportamento linguístico de falantes reais, proporcionando informações altamente confiáveis e isentas de opiniões e de julgamentos prévios sobre os fatos de uma língua. (Trask, 2004)

Corpus na pesquisa linguística:

Por meio de corpus, podem-se observar aspectos morfológicos, morfossintáticos, sintáticos, semânticos, discursivos, etc. bastante relevantes para uma pesquisa linguística.

Pode-se ainda explicar a produtividade e o emprego de palavras, expressões e formas gramaticais.

(BERBER SARDINHA, 2000)

Corpus na pesquisa linguística:

É possível descobrir fatos novos na língua, não perceptíveis pela intuição (BERBER SARDINHA, 2000).

Em resumo, por meio de corpus, descreve-se a língua de forma objetiva.

Sobre corpus

AHDS Guides to Good Practice

Developing Linguistic Corpora: a Guide to Good Practice

Edited by Martin Wynne

Produced by ahds literature, languages and linguistics

ISSN 1463 5194 - Citation Information - Purchasing a Hard Copy (link to Oxbow Books)

http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

Alguns *corpora* disponíveis na web para a pesquisa

- ***Web***
- **Corpus da Folha** (UOL)
- **Lácio-Web**: http://www.nilc.icmc.usp.br/lacioweb/
- PROJETO **COMET** (**CO**rpus **M**ultilíngue para **E**nsino e **T**radução): http://www.fflch.usp.br/dlm/comet/
- **Portal de Corpus – Projeto PLB-BR**: http://www.nilc.icmc.usp.br:8180/portal/news.jsp?id=6

Alguns corpora disponíveis na web para a pesquisa

- **http://www.linguateca.pt/**
  - **CETEMPúblico** (Corpus de Extractos de Textos Electrónicos MCT/Público): corpus de aproximadamente 180 milhões de palavras em português de Portugal
  - **CETENFolha** (Corpus de Extractos de Textos Electrónicos NILC/Folha de São Paulo): *corpus* de cerca de 24 milhões de palavras em português brasileiro com base nos textos do jornal Folha de S. Paulo que fazem parte do corpus NILC/São Carlos.

Alguns corpora disponíveis na web para a pesquisa

- **http://www.linguateca.pt/**
  - **COMPARA:** corpus paralelo de textos em português e inglês
  - **Corpógrafo:** sistema para facilitar a criação de corpora especializados próprios, do tipo «faça-você-mesmo», com capacidades de extração de terminologia.

www.corpusdoportugues.org

O CORPUS DO PORTUGUÊS

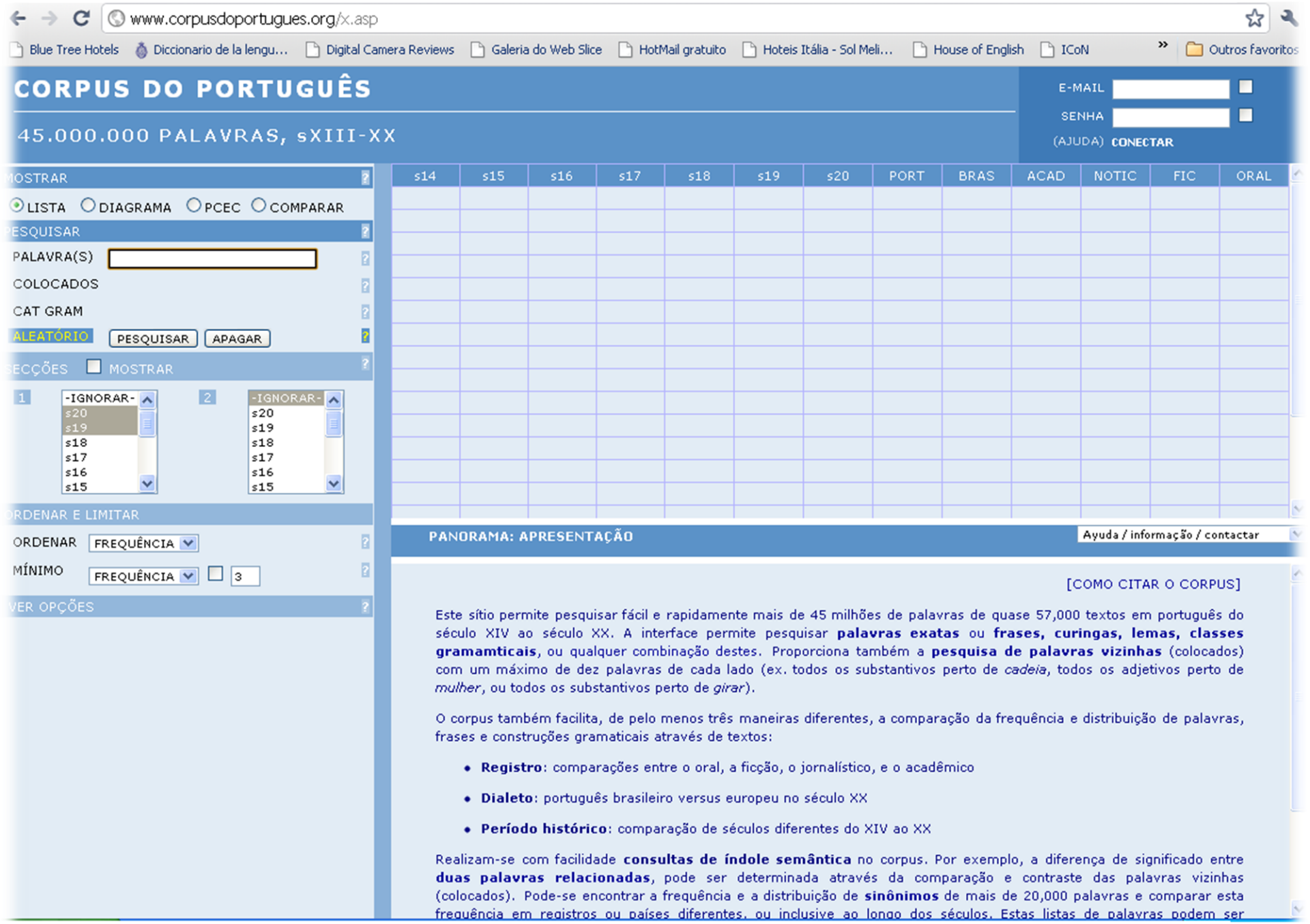
45,000,000 WORDS / PALAVRAS
1300s-1900s

MARK DAVIES
BYU

MICHAEL J. FERREIRA
GEORGETOWN UNIVERSITY

NEH
UNITED STATES
NATIONAL ENDOWMENT
FOR THE HUMANITIES

ENGLISH PORTUGUÊS


www.corpusdoportugues.org/x.asp
Blue Tree Hotels
Diccionario de la lengu...
Digital Camera Reviews
Galeria do Web Slice
HotMail gratuito
Hoteis Itália - Sol Meli...
House of English
ICoN
Outros favoritos
CORPUS DO PORTUGUÊS
45.000.000 PALAVRAS, sXIII-XX
E-MAIL
SENHA
(AJUDA) CONECTAR
MOSTRAR
LISTA DIAGRAMA PCEC COMPARAR
PESQUISAR
PALAVRA(S)
COLOCADOS
CAT GRAM
ALEATÓRIO
PESQUISAR
APAGAR
SECÇÕES MOSTRAR
1
-IGNORAR-
s20
s19
s18
s17
s16
s15
2
-IGNORAR-
s20
s19
s18
s17
s16
s15
ORDENAR E LIMITAR
ORDENAR FREQUÊNCIA
MÍNIMO FREQUÊNCIA 3
VER OPÇÕES
s14 s15 s16 s17 s18 s19 s20 PORT BRAS ACAD NOTIC FIC ORAL
PANORAMA: APRESENTAÇÃO
Ayuda / informação / contactar
[COMO CITAR O CORPUS]
Este sítio permite pesquisar fácil e rapidamente mais de 45 milhões de palavras de quase 57,000 textos em português do século XIV ao século XX. A interface permite pesquisar palavras exatas ou frases, curingas, lemas, classes gramamticais, ou qualquer combinação destes. Proporciona também a pesquisa de palavras vizinhas (colocados) com um máximo de dez palavras de cada lado (ex. todos os substantivos perto de cadeia, todos os adjetivos perto de mulher, ou todos os substantivos perto de girar).
O corpus também facilita, de pelo menos três maneiras diferentes, a comparação da frequência e distribuição de palavras, frases e construções gramaticais através de textos:
Registro: comparações entre o oral, a ficção, o jornalístico, e o acadêmico
Dialeto: português brasileiro versus europeu no século XX
Período histórico: comparação de séculos diferentes do XIV ao XX
Realizam-se com facilidade consultas de índole semântica no corpus. Por exemplo, a diferença de significado entre duas palavras relacionadas, pode ser determinada através da comparação e contraste das palavras vizinhas (colocados). Pode-se encontrar a frequência e a distribuição de sinônimos de mais de 20,000 palavras e comparar esta frequência em registros ou países diferentes, ou inclusive ao longo dos séculos. Estas listas de palavras podem ser

Referências bibliográficas

ALMEIDA, G. M. B.; CORREIA, M. Terminologia e corpus: relações, métodos e recursos. In: Stella E. O. Tagnin; Oto Araújo Vale. (Org.). **Avanços da Lingüística de Corpus no Brasil**. 1 ed. São Paulo: Humanitas/FFLCH/USP, 2008, v. 1, p. 67-94.

ALMEIDA, G.M.B.;VALE, O.A. Do texto ao termo: interação entre Terminologia, Morfologia e Linguística de corpus na extração semiautomática de termos. In: ISQUIERDO, A.N. e FINATTO, M.J.B. (org.) **As ciências do léxico: lexicologia, lexicografia, terminologia** - volume IV. Campo Grande: Ed.UFMS, Porto Alegre: Ed.UFRGS, 2010. p.483-499

ALUÍSIO, S.M.; ALMEIDA, G. M. B. O que é e como se constrói um Corpus? Lições aprendidas na compilação de vários corpora para pesquisa lingüística. **Calidoscópio** (UNISINOS), v. 4, p. 156-178, 2006. Disponível em: http://www.unisinos.br/publicacoes_cientificas/images/stories/pdfs_calidoscopio/vol4n3/art04_aluisio.pdf

Referências bibliográficas

BERBER SARDINHA, T. **Lingüística de corpus**. São Paulo: Manole, 2004.

BERBER SARDINHA, T. Lingüística de Corpus: histórico e problemática. **DELTA,** São Paulo, v. 16, n. 2, 2000.

BERBER SARDINHA, T. Tamanho de corpus. **the ESPecialist**, São Paulo, vol. 23, nº 2. p. 103-122, 2003.

COLETI, J. S.; MATTOS, D. F. ; GENOVES JR., L. C. ; CANDIDO JR., A. ; DI FELIPPO, A. ; ALMEIDA, G. M. B ; ALUÍSIO, S. M. ; OLIVEIRA JR., O.N. A compilação de corpus em língua portuguesa na área de nanociência/nanotecnologia: problemas e soluções. In: Stella E. O. Tagnin; Oto Araújo Vale. (Org.). **Avanços da Lingüística de Corpus no Brasil**. 1 ed. São Paulo: Humanitas, 2008, p. 167-191.

DUBOIS, J; GIACOMO, M.; GUESPIN, L.; MARCELLESI, C.; MARCELLESI, J.B. et MEVEL, J.P. **Dicionário de lingüística**. São Paulo: Cultrix, 1993.

Referências bibliográficas

DUCROT, O. & TODOROV, T. **Dicionário enciclopédico das ciências da linguagem**. 3a ed. São Paulo. Perspectiva. 1998.

GALISSON, R. & COSTE, D. **Dicionário de didáctica das línguas.** Coimbra: Livraria Almedina, 1983.

KILGARRIFF, A. e GREFENSTETTE, G. 2003. Introduction to the Special Issue on Web as Corpus. **Computational Linguistics**, 29(3).

McENERY, T. e WILSON, A. 1996. **Corpus linguistics**. Edinburgh, Edinburgh University Press.

MUNIZ, M. C. M. **A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB.** Dissertação de Mestrado. Instituto de Ciências Matemáticas de São Carlos, USP. 72p. 2004.

Referências bibliográficas

MURAKAWA, C.A.A. 2001. Tradição lexicográfica em língua portuguesa. In: A.M.P.P. OLIVEIRA e A.N. ISQUERDO (orgs.). **As ciências do léxico: lexicologia, lexicografia e terminologia**. 2ª. ed., Campo Grande, Ed. UFMS, p. 153-159.

MURAKAWA, C.A.A. 2006. **António de Morais Silva: lexicógrafo da língua portuguesa**. Araraquara, Laboratório Editorial FCL/UNESP; São Paulo, Cultura Acadêmica Editora, 228 p.

PAUMIER, S. **Unitex user manual**. Disponível em: http://www-igm.univ-mlv.fr/~unitex. 2002.

SINCLAIR, J. 2005. Corpus and Text - Basic Principles. *In*: M. WYNNE (ed.), **Developing Linguistic Corpora: a Guide to Good Practice**. Oxford, Oxbow Books, p. 1-16. Disponível em: http://ahds.ac.uk/linguistic-corpora/. Acesso em: 30/10/2006.

TRASK, R. L. **Dicionário de Linguagem e Lingüística**. São Paulo: Contexto, 2004.

OBRIGADA!

gladis@ufscar.br
www.geterm.ufscar.br

Exercício

1. Entrar no site: **http://www.webcorp.org.uk/**
2. Clicar em *Advanced Search Opti*ons
3. *Search Engine*: **Google**
4. *Case Options: **insensitive***
5. *Search term*: **apagão**
6. *Output format*: **HTML tables (KWIC)**
7. *Web Addresses (URLs)*: **show for concordance lines**
8. *Concordance Span:* **10 *word(s) to left and right***
9. *Number of Pages to Retrieve:* **50**
10. *Site Domain / Country:* **.br**
11. *Textual Domain:* ***all***
12. *Word Filter:* **deixar em branco**
13. Exclude e-mail address from match
14. Submit

WebCorp login

https://wse1.webcorp.org.uk/login/selogin.php

WebCorp login

WebCorp login

WebCorp LSE

Choose a corpus to search:

Small French Newspaper	Arrange Access	About
Anglo-norman Correspondence Corpus	Search	About
Charles Dickens Novels	Arrange Access	About
Restoration Plays	Arrange Access	About
Works of James Joyce	Arrange Access	About
Works of Samuel Beckett	Arrange Access	About
Works of Percy Bysshe Shelley	Arrange Access	About

To arrange access to any of the corpora above please email rdues@bcu.ac.uk.

Username:	Gladis
Firstname:	Gladis
Surname:	Almeida
Email:	gladis.mba@gmail.com
Affiliation:	UFSCar
Interests:	

Change Details

Change Password

Leave Feedback

Logout