Sandra Maria Aluísio
sandra@icmc.usp.br

Gladis Maria de Barcellos Almeida
gladis_maria@uol.com.br

O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa lingüística

What is a *corpus* and how to build it? Lessons learned from developing several linguistic *corpora*

RESUMO - As pesquisas baseadas em *corpus* têm tido na última década um amplo desenvolvimento no contexto brasileiro. Nota-se a sua relevância e pertinência nos domínios da Lingüística, da Lingüística Aplicada e da Lingüística Computacional. Em vista disso, uma abordagem surge para sistematizar procedimentos e dar conta desse novo modo de fazer pesquisa. Essa abordagem é a Lingüística de *Corpus* que, auxiliada pelo desenvolvimento de ferramentas computacionais específicas para o tratamento do português brasileiro, pode alcançar um grande desenvolvimento no Brasil. Entretanto, muito do que já se obteve de desenvolvimento em Lingüística de *Corpus* no cenário internacional não se reflete em muitas das pesquisas realizadas no Brasil, uma vez que as práticas mundialmente aceitas ainda não estão aqui sedimentadas, a despeito de haver no país eminentes pesquisadores que desenvolvem extraordinários projetos baseados em *corpus*. Assim, este artigo tem o propósito de discorrer sobre a concepção de *corpus*, os requisitos e procedimentos para a sua elaboração, os *corpora* e ferramentas existentes e disponíveis e, finalmente, apresentar quatro projetos envolvendo *corpus* cuja descrição e detalhamento pode auxiliar outros pesquisadores nessa tarefa.

Palavras-chave: *corpus*; lingüística de *corpus*; processamento de *corpus*.

ABSTRACT - The research based on *corpus* has had in the last decade an ample development in the Brazilian context. Its relevancy is noticed in the Linguistics, Applied Linguistics and Computational Linguistics research areas. The approach of Corpus Linguistics comes out to systematize procedures and to give account of this new way to make research. The development of Brazilian Portuguese natural language processing tools can help Corpus Linguistics to reach a great development in Brazil. However, the advances in Corpus Linguistics in the international scenery have not happened yet in many of the research carried out in Brazil. The reasons for this is that the procedures and concepts world-wide accepted are not still settled here, in spite of having researchers developing extraordinary projects based on corpus in Brazil. Thus, this article has the intention to discuss several definitions of corpus, the requirements and procedures for its elaboration, the available corpora and tools and, finally, to present four projects involving corpus whose description and detailing can assist other researchers in the corpus building and processing.

Key-words: corpus; corpus linguistics; corpus processing.

A corpus is a remarkable thing, not so much because it is a collection of language text, but because of the properties that it acquires if it is well-designed and carefully-constructed.
(Sinclair, 2005)

Concepção de *corpus* para a Lingüística e para a Lingüística de *Corpus*

A utilização de *corpus* sempre foi um recurso empregado em pesquisas lingüísticas. A título de ilustração, podemos citar a utilização de *corpora* em dicionários elaborados durante os séculos XVIII e XIX, como é o caso do *Vocabulário Portuguez e Latino*, elaborado pelo Padre Rafael Bluteau e publicado entre 1712-1728, embora tenha sido concebido e realizado ainda no século XVII (Murakawa, 2006). O Vocabulário de Bluteau, em oito volumes, foi o primeiro dicionário para o qual foi fixado um

corpus (Murakawa, 2001). Esse *corpus* contendo cerca de 406 obras, aproximadamente, com autores dos séculos XV a XVII, foi utilizado como exemplário de uso lingüístico para as palavras que constavam da nomenclatura do dicionário (Murakawa, 2001; 2006). Outro exemplo já no século XIX é o *Diccionario da Lingua Portugueza*, de Atónio de Morais Silva, segunda edição publicada em 1813, o qual também se valeu de um *corpus* (Murakawa, 2006). O que mudou, portanto, é a concepção de *corpus*. Essa mudança de concepção deve-se à Lingüística de *Corpus*, tida por Berber Sardinha (2004) como uma:

> abordagem que se ocupa da coleta e da exploração de *corpora*, ou conjuntos de dados lingüísticos textuais que foram coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por computador (Berber Sardinha, 2004).

Importa, contudo, definir *corpus*. Há, pelo menos, duas grandes perspectivas a partir das quais se pode definir *corpus*, uma da Lingüística, outra da Lingüística de *Corpus*.

Apresentaremos, a seguir, quatro definições de *corpus* na perspectiva da Lingüística, retiradas de dicionários de Lingüística ou de Linguagem. Para Galisson e Coste (1983), *corpus* é:

> 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. 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. Um *corpus* é chamado exaustivo quando compreende todos os enunciados característicos. E é chamado selectivo quando compreende apenas uma parte desses enunciados.

Para Dubois *et al.* (1993), *corpus* é considerado o conjunto de enunciados a partir do qual se estabelece a gramática descritiva de uma língua. Os autores ainda complementam:

> [o] *corpus* não pode ser considerado como constituindo a língua, mas somente como uma amostra da língua. (...) 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 (...). 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 lingüista deve, pois, procurar obter um *corpus* realmente significativo. Enfim, o lingüista deve desconfiar de tudo o que pode tornar o seu *corpus* não-representativo (método de pesquisa escolhido, anomalia que constitui a intrusão de lingüista, preconceito sobre a língua).

Na concepção de Ducrot e Todorov (2001), *corpus* é um "conjunto, tão variado quanto possível, de enunciados efetivamente emitidos por usuários da referida língua em determinada época". Para Trask (2004), *corpus* é "um conjunto de textos escritos ou falados numa língua, disponível para análise".

Segundo Sinclair, o maior lingüista de *corpus* da história e responsável pelo trabalho pioneiro na área de léxico com o dicionário COBUILD, o primeiro a ser compilado a partir de um *corpus* computadorizado, propõe a seguinte definição para *corpus* na perspectiva da Lingüística 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[4]) [grifo nosso].

Ao observar essas definições, podemos perceber que uma das diferenças entre a concepção da Lingüística de *Corpus* e da Lingüística é o formato do *corpus*, ou seja, os dados devem estar em formato eletrônico. O que significa dizer que uma grande quantidade de livros, ou de revistas, ou mesmo de textos impressos não é considerada *corpus* pela Lingüística de *Corpus*, já que os dados lingüísticos não estão num formato que possam ser processados por computador.

Para outros dois eminentes lingüistas 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:

a) *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;

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

c) *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, diferentemente da idéia que

[1] 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.

se tinha de *corpus* no passado, a qual se referia somente a textos impressos. Ainda de acordo com McEnery e Wilson (1996), o formato possui vantagens consideráveis: i) os *corpora* podem ser pesquisados e manipulados de forma mais rápida; ii) os *corpora* podem ser mais facilmente enriquecidos com informação extra;

d) *referência padrão* (*standard reference*): ainda 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*.

Dentre essas quatro características apontadas pelos autores, a última é digna de nota, já que é uma outra diferença marcante entre a concepção de *corpus* para a Lingüística e para a Lingüística de *Corpus*. Entende-se que disponibilização de *corpus* compilado para futuras pesquisas é uma característica inerente ao *corpus*, de forma que todo o esforço empreendido para a sua construção não seja útil apenas para uma pesquisa, uma vez que se tem uma referência padrão de língua ou de variedade de língua que pode ser utilizada por outros pesquisadores.

Percebe-se, pois, que os dois grandes pontos que diferem entre a Lingüística e a Lingüística de *Corpus* são: o formato computadorizado do *corpus* e a sua posterior disponibilização para outras pesquisas.

Se a Lingüística de *Corpus* descarta livros, revistas e outros textos impressos considerados *corpus* pela Lingüística (pois não estão em formato computadorizado), ela (a Lingüística de *Corpus*) também descarta a Web como *corpus*, ainda que os textos estejam disponíveis e em formato eletrônico, pelo fato de suas dimensões serem desconhecidas, de estar continuamente mudando e pelo fato de não ter sido projetada a partir de uma perspectiva lingüística. Entretanto, é a própria Web que vai facilitar a distribuição e livre acesso de vários *corpora* criados em vários projetos, reforçando uma das características de *corpus* citadas por McEnery e Wilson (1996). Ainda com relação a Web, vale assinalar que existem autores que a consideram um *corpus*, é o caso de Kilgarriff e Grefenstette (2003).

Com relação ao formato computadorizado, é preciso admitir que o surgimento do computador (sobretudo do computador pessoal) interferiu diretamente não só na concepção que se tem de *corpus* como também na sua forma de armazenamento e exploração, já que os recursos oferecidos pelo computador permitiram que uma quantidade antes inimaginável de textos pudesse ser processada na tela em questão de segundos, fazendo com que muitas hipóteses sobre determinados fenômenos lingüísticos pudessem ser testadas rápida e eficientemente. Essa nova forma de armazenamento de textos permitiu a observação e descrição de fenômenos lingüísticos recorrentes antes impossível de perceber, dado que os procedimentos de observação e descrição contavam apenas com recursos manuais.

Sobretudo a partir da década de 1990, os *corpora* passam a ter papel fundamental nas pesquisas lingüísticas, pois data dessa época o início das contribuições advindas da Computação e da Lingüística Computacional. Destacam-se, principalmente, o aprimoramento e desenvolvimento de ferramentas computacionais voltadas para o processamento de língua natural (PLN) do português do Brasil e o efeito que essas ferramentas tiveram para o processamento de *corpus*.

De acordo com Trask (2004), "a partir de *corpora*, podem-se fazer observações precisas sobre o real comportamento lingüístico de falantes reais, proporcionando informações altamente confiáveis e isentas de opiniões e de julgamentos prévios sobre os fatos de uma língua".

Desta forma, por meio de *corpus*, podem-se observar aspectos morfológicos, sintáticos, semânticos, discursivos, etc. bastante relevantes para uma pesquisa lingüística. Podem-se ainda explicar a produtividade e o emprego de palavras, expressões e formas gramaticais. 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.

Questões importantes para o projeto de um *corpus* computadorizado

Para o projeto de um *corpus* computadorizado, devem-se observar um conjunto de requisitos que impactarão na validade e confiabilidade da pesquisa baseada no *corpus*, incluindo se o *corpus* de estudo serve ao propósito inicial da pesquisa (Kennedy, 1998; Biber *et al.*, 1998; Renouf, 1998; Sinclair, 2005): autenticidade, representatividade, balanceamento, amostragem, diversidade e tamanho, os quais serão descritos a seguir.

1) Os textos devem ser *autênticos*. Por autenticidade, compreende-se: a) os textos devem ter sido escritos em linguagem natural, não podendo ser textos "produzidos com o propósito de serem alvo de pesquisa lingüística" (Berber Sardinha, 2000); b) 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).
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 lingüísticas da comunidade cuja língua está

sob análise (Sinclair, 2005). Daí a importância de se fazerem escolhas adequadas, de modo que o *corpus* possa de fato espelhar comportamentos lingüísticos. 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 lingüísticos de uma comunidade?

3) Apesar de Sinclair (2005) afirmar que o conceito de balanceamento é ainda mais vago que o de representatividade, é preciso ter em mente que o *corpus* deve ser *balanceado*, ou seja, deve ter um equilíbrio de gêneros discursivos (informativo, científico, religioso, etc.), ou de tipos de textos (artigo, editorial, entrevista, dissertação, carta, etc.), ou de títulos, ou de autores, ou de todos esses itens juntos, desde que as escolhas sejam adequadas à pesquisa que se pretende realizar, demonstrando que os textos foram escolhidos criteriosamente. Podemos dar como exemplo uma pesquisa que tem por objeto a descrição do pronome de tratamento alocutivo (=você). Uma pesquisa como essa deve, necessariamente, selecionar para o *corpus* o gênero epistolar (composto de cartas), já que é nesse gênero discursivo que pode haver ocorrência significativa do pronome *você*. O mesmo não ocorreria se o gênero escolhido fosse o jornalístico, por exemplo.
4) Biber *et al*. (1998) advoga que uma *amostragem* proporcional não é adequada para *corpus* de língua, pois esta deveria ser organizada demograficamente. Entretanto, tal tipo de *corpus* não representaria os tipos de gêneros e de textos, pois um *corpus* com tal amostragem poderia conter 90% de conversação, 3% de cartas e notas e 7% divididos entre tipos de textos tais como reportagens e notícias, revistas, artigos acadêmicos, literatura, aulas, e escrita não publicada, pois são poucas as pessoas que publicam ou mesmo falam para uma grande audiência. Para o estudo da língua importa um *corpus* com amostras que sejam representativas por incluírem toda a variação lingüística que existe.
5) Com relação à *diversidade*, Biber *et al*. (1998) enfatiza que não existe o que chamamos de "língua geral", dado que cada gênero e tipo de texto têm seus próprios padrões de uso. Desta forma, se um *corpus* se presta para estudos de variação ou procura representar uma língua, ele deve se preocupar com a diversidade de gêneros e tipos de textos, com a variação de dialetos e, por último, com uma diversidade de tópicos que é de fundamental importância para estudos lexicográficos, pois a freqüência de muitas palavras varia de acordo com a variação de tópicos. Este último tipo de diversidade deve ser considerado para todos os tipos de estudos.
6) Segundo Sinclair (2005), o *corpus* deve ter o *tamanho* adequado ao tipo de pesquisa que se vai realizar e à metodologia a ser adotada na pesquisa. Quando se fala em tamanho de um *corpus*, não se trata somente do número total de palavras (*tokens*) e de palavras diferentes (*types*), mas com quantas categorias (gêneros discursivos, tipos de textos, datas, autores, etc.) um *corpus* deve contar, quantas amostras de cada categoria e quantas palavras existem dentre de cada amostra (Kennedy, 1998). Para estudos da prosódia, por exemplo, um *corpus* de 100 mil palavras será o suficiente para generalizações com propósitos descritivos; para estudos de muitos processos sintáticos, um *corpus* de 500 mil a 1 milhão de palavras é suficiente; para a criação de dicionários de língua geral, que devem definir os vários significados de suas entradas, gramáticas e usos, seria necessário um *corpus* muito maior, por exemplo, o *Bank of English*[2] que apóia a criação de produtos da editora Collins possui atualmente 530 milhões de palavras.

Para Biber (1993), a elaboração de um *corpus* é um processo que avança em ciclos: inicia-se a escolha de textos baseada em critérios externos culturalmente aceitos (tipologia de gêneros e tipos de textos, por exemplo), depois se prossegue com investigações empíricas da língua ou variedade lingüística sob análise (também denominados critérios internos) e, finalmente, procede-se com a revisão de todo o projeto.

Etapas metodológicas para a compilação de um *corpus*

Embora existam muitos *corpora* disponíveis tanto livremente como mediante pagamento (as taxas geralmente são modestas para pesquisa acadêmica) – a partir dos quais se pode gerar um sub*corpus* de estudo ou mesmo tomar o *corpus* todo como uma unidade, dependendo da questão de pesquisa[3] –, ainda pode ser necessário compilar um *corpus* próprio. Para a compilação de tal *corpus*, existem três estágios principais a seguir: 1) projeto do

[2] http://www.titania.bham.ac.uk/.

[3] Por exemplo, estudo de um autor em particular, o qual não se encontra representado em algum corpus, ou de um gênero mais atual como os e-mails e chats, estudo de textos de épocas não cobertas pelos corpora ou ainda estudo de um fenômeno raro.

corpus, que inclui a seleção dos textos e os cuidados com os requisitos que foram discutidos na seção anterior, 2) compilação (ou captura), manipulação, nomeação dos arquivos de textos, e pedidos de permissão de uso, e 3) anotação.

Projeto de corpus: a seleção dos textos

Inicialmente, procede-se à seleção dos textos pertinentes e relevantes para a pesquisa. Para esta etapa, a definição do tipo de *corpus* que está se compilando é importante; outras decisões dizem respeito ao seu tamanho e à sua composição em termos dos textos existentes bem como dos gêneros aos quais eles pertencem.

Existem várias tipologias de *corpus* que indicam os parâmetros importantes de consideração. Uma das mais antigas é a de Atkins *et al.* (1992) e uma bastante atual é a de Berber Sardinha (2004) que inclui sete critérios. Dentre eles, o mais importantes é o critério *modalidade* (texto falado, escrito ou ambos) e suas proporções (dado que a compilação de um *corpus* de fala é bastante cara).

Compilação e manipulação do corpus

A **compilação** consiste no armazenamento em arquivos predeterminados de todos os textos selecionados.

Podem-se buscar textos provenientes da Web ou mesmo textos impressos, nesse caso, será necessário digitalizá-los e corrigir o resultado do processo de OCR (*optical character recognition*) devido a erros comuns durante o reconhecimento de caracteres, mesmo existindo atualmente bons produtos.

Para o caso de se utilizar a Web, especificamente, existem duas grandes opções na obtenção de textos, as quais se subdividem como segue:

1) a busca na Web com máquinas de busca:
   a. uso de uma máquina de busca como o *Google* para pesquisar toda a Web (podem-se utilizar palavras-chave escolhidas para a pesquisa em foco, sobretudo no caso de pesquisas terminológicas);
   b. uso de ferramentas que pré-processam e/ou pós-processam os resultados das buscas de tais máquinas como fazem o *WebCorp*[4] e *KWiCFinder*[5];
2) a coleta de páginas da Web, organizando-as num computador local:
   a. construção automática de *corpus* com ajuda de *offline browsers* como o HTTrack[6] ou com ajuda de ferramentas de apoio para a compilação de *corpora* descartáveis (*disposable corpora*) como o Corpógrafo[7] e o *Toolkit BootCat*[8], os quais geralmente realizam limpeza de tabelas, referências, agradecimentos, etc. e/ou revisão ortográfica se essa operação for importante para a pesquisa (por exemplo pesquisa terminológica);
   b. coleta do *corpus* pela seleção de páginas de forma manual ou semi-automática de acordo com um projeto específico de *corpus*. Esta última opção não é diferente da forma como grandes *corpora*, como o BNC[9], foram construídos.

A *manipulação* do *corpus* compõe-se das seguintes atividades:

a) conversão manual e automática (por exemplo, com o pacote XPDF[10]) de formatos "doc", "html" e "pdf" para "txt";
b) limpeza e formatação, de maneira a preparar o *corpus* para o processamento computacional, o que significa tirar imagens, gráficos, tabelas, números de páginas e demais anotações que não fazem parte do texto propriamente dito. A limpeza e a formatação possibilitam o processamento do *corpus* por ferramentas computacionais, como por exemplo contador de freqüência, concordanciador, ferramenta de extração automática de termos, etc.

Nomeação de arquivos e geração de cabeçalhos

Depois que todos os textos forem convertidos em formato "txt", eles devem receber um nome. Ressalte-se que essa nomeação deve seguir determinado padrão de forma a facilitar a recuperação posterior de cada texto.

Proteção da identidade dos participantes de um corpus e pedidos de direitos de uso dos textos

Na compilação de *corpus*, devem-se seguir as regras legais para obtenção de direitos de uso do material junto a autores e editores que detêm o *copyright* do texto ou consentimento de indivíduos cujos direitos de privacidade devem ser reconhecidos. Esta é uma etapa da compilação de um *corpus* que não é técnica, é demorada e tediosa, marcada por inúmeras negociações que podem se

[4] http://www.webcorp.org.uk/.
[5] http://miniappolis.com/KWiCFinder/KWiCFinderHome.html.
[6] http://www.httrack.com/.
[7] http://poloclup.linguateca.pt/corpografo/.
[8] http://sslmit.unibo.it/~baroni/bootcat.html.
[9] http://www.natcorp.ox.ac.uk/.
[10] XPDF é um programa de código aberto que permite a conversão automática de arquivos, conferir: http://www.foolabs.com/xpdf/.

arrastar por anos – muitas vezes esta é a razão de muitos *corpora* simplesmente não estarem disponíveis publicamente. Uma estratégia importante para vencer a negação do pedido de permissão de uso é a coleta de um número maior de textos dentro de cada categoria de um *corpus* (gênero, tipos de textos, data) para se preparar para o caso da permissão não ser concedida.

Em Hasund (1998), discute-se como foi realizada a proteção da identidade dos participantes do *corpus* COLT (*The Bergen Corpus of London Teenage Language*), um *corpus* de 500 mil palavras de língua falada coletado em 1993 na University of Bergen, Noruega. Na versão transcrita do COLT (e na parte correspondente do BNC), sobrenomes, endereços, números de telefones foram removidos, embora os nomes sejam reais, isto é, não foram trocados por fictícios.

Não existe, entretanto, nenhuma abordagem amplamente aceita para preservar o anonimato de indivíduos em *corpus* da modalidade oral. A tendência é pelo completo anonimato, ou seja, apagamento de nomes, sobrenomes e títulos profissionais, nomes de animais de estimação, endereços e telefones (que são removidos ou trocados por códigos). Um outro procedimento é a troca de nomes por similares equivalente prosodicamente aos originais. No *corpus Bank of English*, por exemplo, todos os nomes foram trocados por códigos indicando o gênero do falante, mais um número que corresponde a uma descrição de cada um mantida separadamente.

Enquanto aspectos éticos e legais da preservação do anonimato tratam do interesse do informante, aspectos sociolingüísticos e computacionais tratam do interesse da pesquisa. Nomes e apelidos, por exemplo, fornecem informações sociolingüísticas relacionadas à característica socioeconômica e grupo étnico, entretanto, fazer a troca por outro nome que preencha todos os critérios sociolingüísticos consome muito tempo, razão pela qual raramente é feita.

Anotação

Em relação à anotação, são dois basicamente os níveis de representação das informações presentes num *corpus*: a anotação estrutural e a anotação lingüística.

A *anotação estrutural* compreende a marcação de dados externos e internos dos textos. Como dados externos entendemos a documentação do *corpus* na forma de um cabeçalho que inclui os metadados textuais (ou dados estruturados sobre dados), isto é, dados bibliográficos comuns, dados de catalogação como tamanho do arquivo, tipo da autoria, a tipologia textual e informação sobre a distribuição do *corpus*. Como dados internos temos a 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, nomes, referências, datas e ênfases tipográficas do tipo negrito, itálico, sublinhado, etc.

Essas informações de cabeçalho facilitam a posterior recuperação do texto bem como a geração de sub*corpus*, isto é, podem-se selecionar todos os textos de determinado autor, ou de determinada época, ou de determinado gênero, etc.

A *anotação lingüística* pode ser em qualquer nível que se queira, isto é, nos níveis morfossintático, sintático, semântico, discursivo, etc., sendo inserida de três formas: manualmente (por lingüistas), automaticamente (por ferramentas de Processamento de Língua Natural – PLN) ou semi-automaticamente (correção manual da saída de outras ferramentas). Essa última é comprovadamente mais eficiente, pois revisar é mais rápido e gera dados mais corretos do que anotar pela primeira vez.

Um padrão que vem sendo usado atualmente para anotação de corpus para a criação de aplicações de PLN é o XCES[11] (Corpus Encoding Standard for XML) que foi derivado do TEI[12] (Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange), este último largamente usado para criação de corpora contemporâneos ou históricos, para pesquisas terminográficas ou lexicográficas, estudos literários, descrições lingüísticas, entre outras.

Alguns corpora *disponíveis na Web para pesquisa*

Como a construção de um *corpus* nos moldes acima mencionados não é tarefa simples e rápida, antes de construir um, é útil saber se *corpus* com determinadas características já existem. Nesse sentido, apresentaremos alguns *corpora* disponíveis na Web, os quais podem ser utilizados para muitas pesquisas.

Arquivos da Folha *(http://www1.folha.uol.com.br/folha/arquivos/):*

Está disponível na Web o texto integral de todas as edições do jornal desde 1994. Todo esse material é extremamente útil para fazer buscas por conteúdo ou mesmo para atestar freqüência e emprego de determinadas palavras ou expressões na língua, no gênero jornalístico. É um *corpus* muito rico, entretanto, tem alguns inconvenientes: a) é acessível somente para assinantes do jornal *Folha de S. Paulo* ou do *Universo On Line (UOL)*; b) a busca ocorre ano a ano, isto é, não é possível conferir, por exemplo, a freqüência de uma expressão em todos os anos, mas deve-se selecionar o ano e digitar a expressão que se deseja

[11] http://www.cs.vassar.edu/XCES/
[12] http://etext.lib.virginia.edu/standards/tei/teip4/index.html

pesquisar; c) a expressão pesquisada não aparece na tela no formato de um concordanciador, mas pequenos contextos com *links* são oferecidos ao usuário, de forma que, acionado esses *links*, é possível chegar aos textos na íntegra; d) não é possível gerar sub*corpus*, isto é, selecionar as edições desejadas e fazer *download*, todas as buscas são feitas de forma *on-line* no *site* da *Folha*.

Lácio-Web *(http://www.nilc.icmc.usp.br/lacioweb/):*

O Lácio-Web[13] (LW) é um projeto organizado pelo Núcleo Interinstitucional de Lingüística Computacional (NILC[14]), em parceria com o Instituto de Matemática e Estatística (IME) e a Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH), todos pertencentes à Universidade de São Paulo (USP). O LW disponibiliza livremente na Web: a) vários *corpora* do português brasileiro escrito contemporâneo, representando bancos de textos adequadamente compilados, catalogados e codificados em um padrão que possibilite fácil intercâmbio, navegação e análise; e b) ferramentas lingüístico-computacionais, tais como contadores de freqüência, concordanciadores e etiquetadores morfossintáticos.

Projeto COMET *(*Corpus *Multilíngüe para Ensino e Tradução – http://www.fflch.usp.br/dlm/comet/:*

O projeto COMET, em elaboração junto ao Centro Interdepartamental de Tradução e Terminologia (CITRAT) da Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH) da USP, disponibiliza um *corpus* eletrônico que tem por objetivo servir de suporte a pesquisas lingüísticas, principalmente nas áreas de tradução, terminologia e ensino de línguas. O COMET é composto por três sub*corpora*: a) *Corpus* Técnico-Científico – CorTec: *corpus* comparável de textos técnicos e/ou científicos originalmente escritos em português brasileiro e em inglês; b) *Corpus* Multilíngüe de Aprendizes – CoMAprend: constituído de redações dos alunos da graduação e dos cursos de extensão das áreas do Departamento de Letras Modernas: alemão, espanhol, francês, inglês e italiano; c) *Corpus* de Tradução – CorTrad: subdivide-se em Literário e Juramentado; o *corpus* Literário é composto de contos traduzidos do inglês e seus respectivos originais, o *corpus* Juramentado será constituído de textos cedidos pela Junta Comercial de São Paulo por meio de contrato de comodato com a USP.

Linguateca *(http://www.linguateca.pt/):*

A Linguateca é um centro de recursos para o processamento computacional da língua portuguesa e tem como objetivo servir à comunidade que se dedica ao processamento do português. No *site* da Linguateca estão disponíveis, entre outros, os seguintes *corpora* crus e anotados pelo analisador sintático *Palavras*[15]: a) CETEMPúblico (*Corpus* de Extratos de Textos Eletrônicos MCT/Público – http://www.linguateca.pt/CETEMPublico/): *corpus* de aproximadamente 180 milhões de palavras em português europeu, criado pelo projeto *Processamento computacional do português* (projeto que deu origem à Linguateca) após a assinatura de um protocolo entre o Ministério da Ciência e da Tecnologia (MCT) português e o jornal *Público* (jornal português) em abril de 2000; b) CETENFolha (*Corpus* de Extractos de Textos Electrónicos NILC/Folha de São Paulo – http://www.linguateca.pt/CETEMPublico/): *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; c) COMPARA (http://www.linguateca.pt/COMPARA/): *corpus* paralelo que tem como base textos em português e as suas traduções para inglês e textos em inglês e as suas traduções para português.

Algumas ferramentas disponíveis na Web

Há disponível gratuitamente na Web uma série de ferramentas que podem auxiliar a pesquisa envolvendo *corpus*. Apresentaremos, inicialmente, as ferramentas de processamento de *corpora* gerais ou especializados, as quais incluem o *WebCorp* e o *Unitex*. Em seguida, as ferramentas de geração e gerenciamento de *corpora* especializados, abrangendo o Corpógrafo e o *ToolKit BootCaT*.

Ferramentas de processamento de corpus

WebCorp

WebCorp é um conjunto de ferramentas que permitem acesso a Web como um recurso lingüístico, isto é, permitem extrair fatos sobre várias línguas como se a Web fosse um *corpus* – o maior deles[16]. Versões *demo* desse conjunto de ferramentas são disponibilizadas gratuitamente na Web a partir do endereço http://www.webcorp.org.uk/. Vale assinalar que está em corrente desenvolvimento a construção de uma máquina de busca lingüística para melhorar o desempenho do *WebCorp*.

---

[13] O projeto Lácio-Web será detalhado a seguir.

[14] Localizado no Instituto de Ciências Matemáticas e de Computação, da Universidade de São Paulo (USP), campus de São Carlos (SP, Brasil), www.nilc.icmc.usp.br/.

[15] Desenvolvido por Eckhard Bick (http://visl.hum.sdu.dk/).

[16] Em nível internacional, houve dois Workshops dedicados ao tema "Web as a corpus" - o primeiro em conjunto com a conferência Corpus Linguistics 2005, e o segundo em conjunto com a 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).

WebCorp pode ser usado por pesquisadores e professores de língua, por exemplo, que tenham interesse em analisar como certas palavras e expressões são usadas, especialmente as palavras raras ou neologismos que não aparecem em dicionários e em *corpora* padrões. Desde seu lançamento, em 2000, pela *Research and Development Unit for English Studies* (RDUES) na *School of English* da *University of Central England,* Birmingham, *Webcorp* tem sido usado por lingüistas, lexicógrafos, alunos e professores de línguas, editores, jornalistas, publicitários e demais pesquisadores provenientes de distintas áreas.

WebCorp possui uma interface similar a muitas máquinas de busca (observe-se a tela principal na Figura 1) na qual se pode digitar uma palavra ou expressão de busca, escolher as opções nos *menus* e clicar o botão "*Submit*". Ele trabalha com os resultados do motor de busca escolhido (há opções para quatro deles: *Google, Altavista, Metacrawler* e *AllTheWeb*), tomando a lista de URLs[17] retornada do motor de busca escolhido e extraindo concordâncias de cada página. Todas as concordâncias são apresentadas em uma única página separadas por arquivo da Web e com *links* para os *sites* de onde vieram (observe-se parte do resultado da palavra "corpus" na Figura 2).

Uma das opções avançadas que merece destaque é a escolha de busca em um dado domínio, por exemplo, ao escolher ".ac.uk" ela será restrita às instituições acadêmicas do Reino Unido; ".fr" às URLs da França e ".br.com" às URLs de empresas no Brasil. Outra opção é a possibilidade de analisar colocações da palavra de busca, isto é, as palavras que aparecem com freqüência maior nas proximidades da palavra em foco, podendo também excluir *stopwords* na apresentação das colocações. A Figura 3 apresenta as colocações da palavra "corpus" em URLs do domínio ".ac.uk", excluindo *stopwords*.

Unitex

O *Unitex* consiste em um conjunto de programas para processamento de *corpus* lingüísticos composto por uma interface gráfica em *Java* e diversos programas desenvolvidos em C (Paumier, 2002). A interface *Java* em conjunto com os programas em C permitem que a ferramenta possa ser portada para uma série de plataformas sem perdas significativas de desempenho durante o processamento de *corpus*.

Dentre os recursos lingüísticos oferecidos estão dicionários[18] e tabelas do léxico-gramática[19]. Os dicionários contêm palavras simples e compostas de um idioma além de informações gramaticais sobre cada palavra. As gramáticas

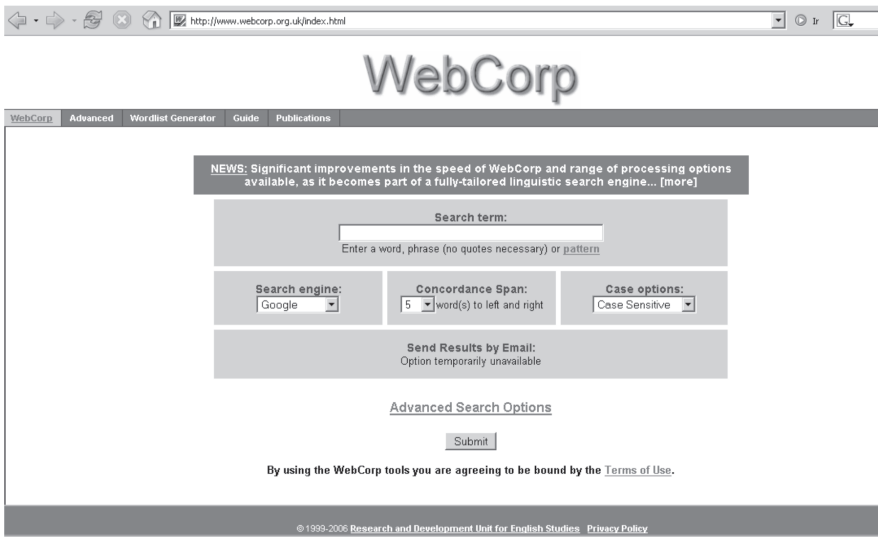


Figura 1. Tela principal do *WebCorp* a partir da qual se podem escolher as opções do *menu* e acessar as opções avançadas de busca.

[17] "Sigla que designa a localização de um objeto na Internet (rede mundial de computadores), segundo determinado padrão de atribuição de endereços em redes." (Novo Dicionário Eletrônico Aurélio versão 5.0, 2004)

[18] Dicionários para serem utilizados pela máquina e não para humanos.

[19] As tabelas do léxico-gramática são matrizes binárias nas quais as linhas são ocupadas por entradas do léxico e nas colunas são explicitadas as propriedades sintático-semânticas de cada entrada lexical. No cruzamento de cada coluna com cada linha são colocados um sinal de '+' no caso da propriedade se aplicar àquela entrada, e um '-' para o caso contrário. Essa metodologia foi proposta por M.Gross (1968, 1975) no estudo dos verbos do francês e tem sido aplicada a diversas línguas no estudo principalmente de elementos predicativos como os verbos, adjetivos e substantivos predicativos. Uma bibliografia a respeito dessa teoria/metodologia pode ser encontrada em: http://ladl.univ-mlv.fr/ (Vale, 1998 e 2001).

são representadas por meio de autômatos de texto, um formalismo baseado em autômatos finitos. As tabelas do léxico-gramática mostram as propriedades de algumas palavras. A versão 1.2 da ferramenta provê suporte para mais de 14 idiomas (incluindo o Português). Entretanto, o usuário pode adicionar facilmente suporte a qualquer idioma graças ao uso do padrão *Unicode*[20] para codificação de texto. O suporte ao idioma português é particularmente bom gra-

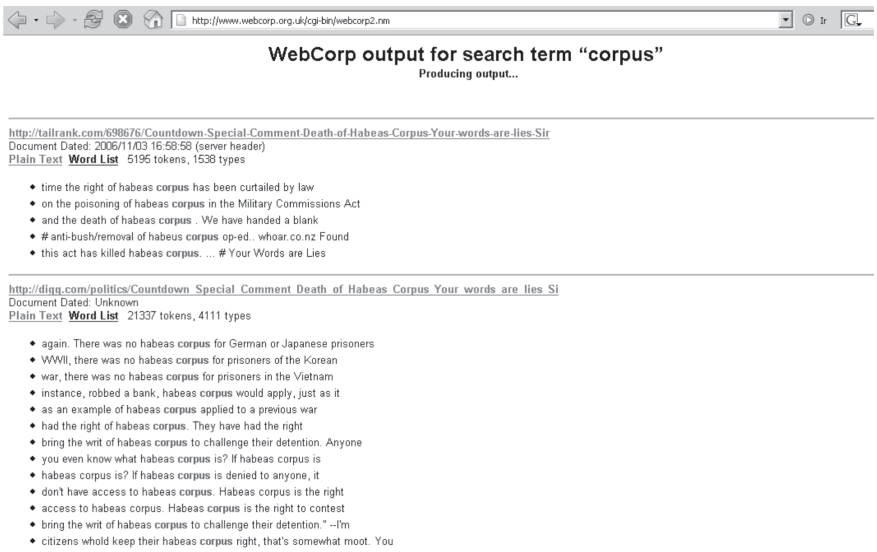
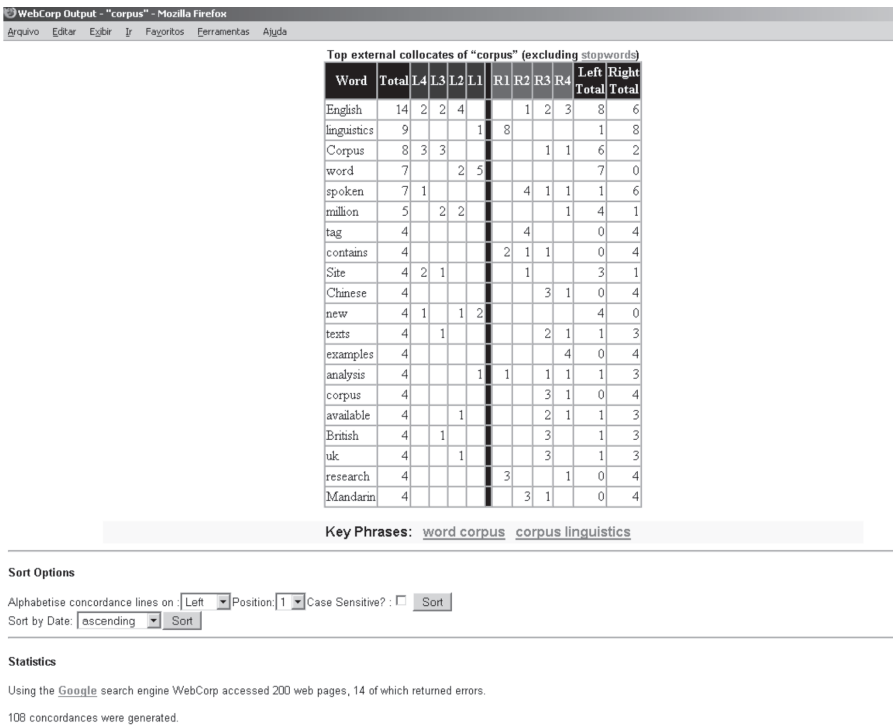


Figura 2. Parte do resultado da busca da palavra "corpus", com as opções de apresentação de 5 palavras à esquerda e à direita da palavra em foco.



Word	Total	L4	L3	L2	L1	R1	R2	R3	R4	Left Total	Right Total
English	14	2	2	4			1	2	3	8	6
linguistics	9				1	8				1	8
Corpus	8	3	3					1	1	6	2
word	7			2	5					7	0
spoken	7	1					4	1	1	1	6
million	5		2	2					1	4	1
tag	4						4			0	4
contains	4					2	1	1		0	4
Site	4	2	1				1			3	1
Chinese	4							3	1	0	4
new	4	1		1	2					4	0
texts	4		1					2	1	1	3
examples	4								4	0	4
analysis	4				1	1		1	1	1	3
corpus	4							3	1	0	4
available	4			1				2	1	1	3
British	4		1					3		1	3
uk	4			1				3		1	3
research	4					3			1	0	4
Mandarin	4						3	1		0	4

Figura 3. Colocações à esquerda e à direita da palavra "corpus" a partir de 200 páginas do domínio ".ac.uk". Expressões padrões selecionadas deste conjunto foram "word corpus" e "Corpus Linguistics" que são apresentadas como links prontos para serem analisados a partir do Google. As colocações estão ordenadas pela freqüência.

[20] http://unicode.org/

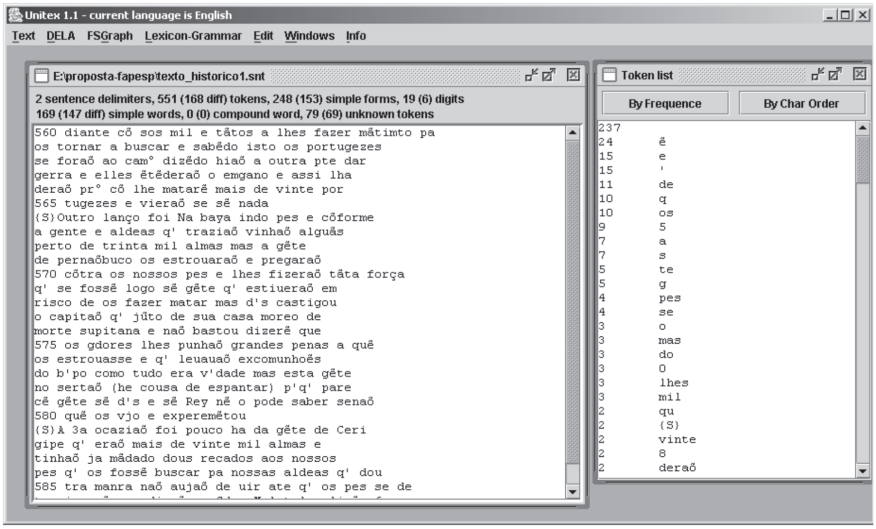


Figura 4. Texto segmentado e lista de *tokens*. À esquerda vemos um texto após a fase de segmentação e pré-processamento; à direita são exibidos os *tokens* extraídos do texto.

ças ao trabalho Unitex-PB desenvolvido em Muniz (2004) e Muniz *et al.* (2005).

O software *Unitex* é uma implementação livre do programa *Intex*, ambos criados no laboratório francês LADL[21] (*Laboratoire d'Automatique Documentaire et Linguistique*), por isso as funcionalidades fornecidas por essas ferramentas são bem semelhantes. Os dicionários *Unitex* se baseiam no formalismo DELA (*Dictionnarie Electronique du LADL*) também desenvolvido no laboratório LADL.

Pré-processador de textos

Um arquivo de texto não formatado (formato "txt") com codificação *Unicode* é convertido para uma forma pré-processada após ser aberto pela primeira vez no *Unitex*. Os arquivos pré-processados geralmente possuem a extensão ".snt". Durante o processo de conversão, o texto original é segmentado em sentenças e unidades lexicais (*tokens*) (Figura 4). Além disso, repetições desnecessárias de caracteres de separação, tais como espaços, quebras de linha e tabulações, são removidas e formas não ambíguas do texto são normalizadas para simplificar operações de busca, sendo que as normalizações são definidas pelo usuário (Figura 5). Como exemplo, a palavra "daí" é normalizada em "de aí". É importante notar que normalização não pode ocorrer para palavras ambíguas tal como a palavra "desse" que pode significar "de esse" ou uma conjugação do verbo "dar".

Nesta etapa, é possível construir um autômato de texto sobre o arquivo de entrada. Além disso, também é possível aplicar um conjunto de dicionários de palavras simples

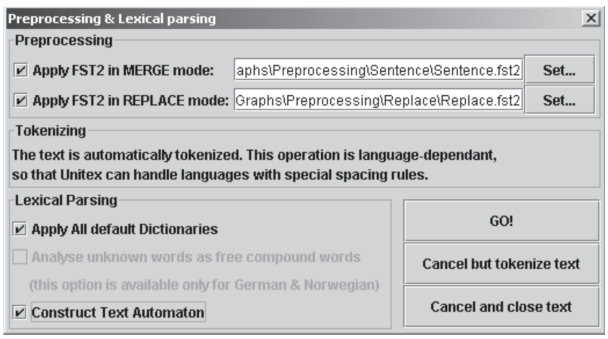


Figura 5. Pré-processador.

e compostas durante o pré-processamento para a construção de um subconjunto de dicionários contendo apenas as palavras presentes no texto. Neste processo, as palavras dos textos são agrupadas em 3 classes: palavras simples, palavras compostas e palavras não reconhecidas (Figura 6). As únicas tarefas apresentadas acima necessárias durante o pré-processamento são a segmentação em unidades lexicais e a remoção de caracteres de separação desnecessários, as demais podem ser efetuadas posteriormente.

Na Figura 5 é exibida a caixa de diálogo para pré-processamento de textos sem formatação. Os textos são segmentados de acordo com as regras definidas no arquivo indicado na opção "*Apply FST2 in MERGE mode*". O arquivo definido em "*Apply FST2 in REPLACE mode*" contém regras de normalização de formas não ambíguas. A opção "*Construct Text Automaton*" permite a criação de autômatos de texto. A opção "*GO!*" inicia o pré-processamento do texto.

[21] http://ladl.univ-mlv.fr/.

A Figura 6 mostra um dicionário (esquerda) onde são listadas informações morfossintáticas das palavras reconhecidas. As palavras estão divididas em três grupos: palavras simples; palavras compostas e palavras não reconhecidas. A direita pode ser observado o autômato de texto para uma sentença pertencente a um texto histórico.

Concordanciador

O concordanciador presente na ferramenta permite a busca de padrões através de expressões regulares. Seqüências de símbolos reservadas são utilizadas para denotar uma expressão regular. As operações de concatenação, união, fecho de *Kleene* e negação são permitidas e representadas respectivamente pelos símbolos: ".", "+", "*", "!". Por exemplo, a expressão regular "para.dizer*" representa a palavra "para" imediatamente seguida por zero ou mais ocorrências da palavra "dizer". As seqüências de símbolos abaixo realizam operações úteis:

* <E>: representa uma cadeia vazia
* <MOT>: qualquer seqüência de letras do alfabeto
* <MIN>: qualquer seqüência de letras minúsculas
* <MAJ>: qualquer seqüência de letras maiúsculas
* <PRE>: uma seqüência de letras começando por maiúsculas
* <NB>: qualquer seqüência de algarismos
* <^>: representa o caractere de quebra de linha
* #: impede a presença de espaço em branco

Adicionalmente, é possível representar nas expressões regulares informações codificadas nos dicionários. Por exemplo, a expressão <A> denota qualquer adjetivo, já a expressão <dizer.V> denota qualquer palavra que tenha "dizer" como sua forma canônica e seja da classe dos verbos. Um exemplo de busca mais avançada pode ser dado pela expressão <V><A> que faz a busca de um verbo seguido de um adjetivo (figura 7).

Dicionários

Existem dois tipos principais de dicionários no formato DELA: os dicionários de forma canônica (DELAS) e os dicionários de formas flexionadas (DELAF). Além disso, existem duas variantes para palavras compostas: DELAC para formas canônicas e DELACF para formas flexionadas. A ordem de prioridade em pesquisas em dicionários é definida pelos símbolos "+" (mais prioritário) e "-" (menos prioritário) adicionados no fim dos nomes de arquivos de cada dicionário.

Uma possível entrada para um dicionário DELAF é dada por "abandonou,abandonar.V:J3s/comentário". Esta entrada indica que a palavra "abandonou" possui a forma canônica "abandonar", sendo "abandonar" um verbo. O itens "J3s" indica terceira pessoa do pretérito, e a seqüência depois do símbolo "/" indica um comentário. Símbolos reservados podem ser representados como parte de uma entrada se forem antecedidos pelo símbolo "\".

O formato das entradas nos demais dicionários é semelhante ao formato do exemplo mostrado acima com pequenas variações. Além disso, é possível armazenar informações semânticas adicionais por meio de palavras reservadas como por exemplo "AnlColl" e "ConcColl". A primeira indica um coletivo de animais (exemplo: manada) enquanto que a segunda indica um coletivo humano (exemplo: banda).

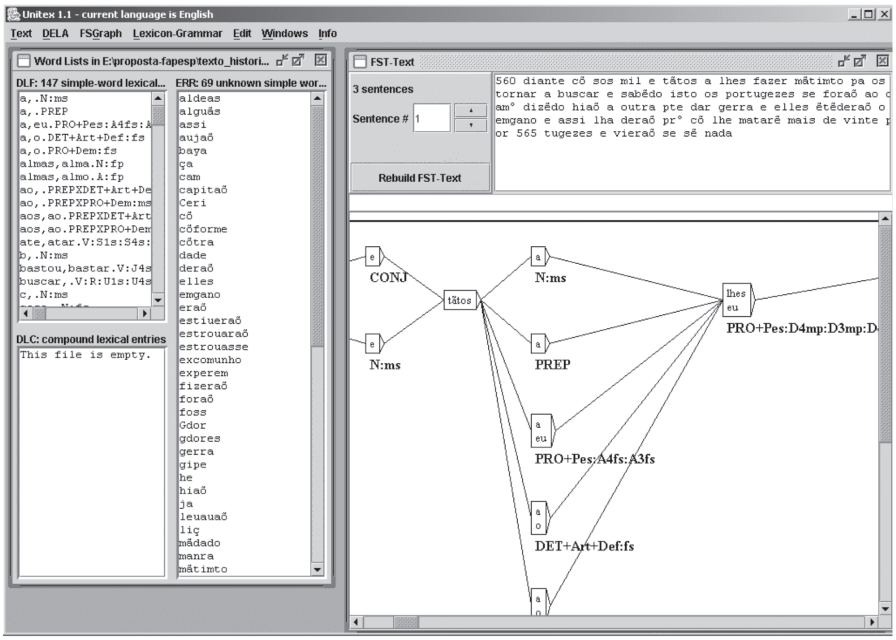


Figura 6. Dicionário morfossintático e grafo de texto

Figura 7. Busca por verbo seguido de adjetivo

O *Unitex* fornece recursos para tratar dicionários no formato DELA. É possível comprimir um dicionário, verificar se contém erros de formatação ou ordená-lo caso ainda não esteja em ordem alfabética. Como os mesmos símbolos podem ser ordenados de maneiras diferentes de acordo com o idioma em uso, o usuário pode definir seus próprios critérios de ordenação por meio de um arquivo chamado "*Alphabet_sort.txt*".

Alguns códigos gramaticais são utilizados para permitir a flexão automática de uma forma canônica. Um novo dicionário contendo as formas flexionadas pode ser gerado automaticamente pelo *Unitex* a partir do dicionário original e de uma gramática de flexão previamente definida.

Ferramentas de geração e gerenciamento de corpora especializados

O Ambiente Corpógrafo

Desenvolvido pela Faculdade de Letras da Universidade do Porto (FLUP), o Corpógrafo[22] é um gestor de *corpus* que se encontra, atualmente, direcionado para pesquisas terminológicas, isto é, a extração de termos e sua organização em bases de dados. Fornece um ambiente Web integrado para o manejo de *corpus*, disponibilizando ferramentas para processamento de *corpus*. Dentre as ferramentas que possui, estão concordanciadores, contadores de freqüência e também ferramentas de pré-processamento de *corpus*, como as de limpeza de *corpus* e sentenciadores. Toda funcionalidade do Corpógrafo está associada a um dos quatro ambientes de trabalho ou módulos: gestor de ficheiros, pesquisa de *corpora*, centro de conhecimento e centro de documentação, essa subdivisão diminui a sobrecarga de trabalho no ambiente.

Dos quatro módulos contidos no Corpógrafo, o que mais interessa para este artigo é o "Gestor de ficheiros", que trata especificamente da montagem de *corpus*. Para construir um *corpus* no Corpógrafo, primeiramente é necessário selecionar os textos que comporão o *corpus*, que podem ser fornecidos de duas maneiras: ou enviando o próprio arquivo (*upload*) ou informando a URL onde o arquivo pode ser encontrado. O Corpógrafo aceita textos do tipo "pdf", html, "doc", "ps" e "rtf", além do "txt", formato para o qual todos os outros tipos de texto são transformados. O Corpógrafo oferece ferramentas para o pré-processamento desses textos, tais como sentenciadores (denominados "fraseadores" em português de Portugal) e um ambiente de edição que permite fazer a "limpeza" de textos (retirar lixo provindo da conversão de tipos de texto, remoção de cabeçalhos, tabelas, referências ou agradecimentos). Após pré-processar os textos, pode-se selecionar aqueles que farão parte do *corpus*.

Tendo um *corpus* montado seguindo os passos anteriores, o Corpógrafo oferece ferramentas de busca e extração de conhecimento de *corpus*, como um concordanciador com suporte para pesquisas utilizando expressões regulares, gerador de n-grama[23] (sendo 5 o tamanho máximo possível para o n-grama), extratores de

[22] http://www.linguateca.pt/Corpografo/

[23] Lexias com número variável de palavras.

terminologia, relações semânticas e mapas conceituais, dentre outras.

O ToolKit BootCaT

O *BootCaT*[24], extrator automático de *corpus* e de termos (do inglês "*Bootstrapping Corpora and Terms*"), propõe a montagem de *corpus*, de modo iterativo, a partir de textos obtidos na Web. O *BootCaT* é composto por várias ferramentas escritas em Perl[25], que foram projetadas para executar pequenas partes do processo de montagem de *corpus*.

Basicamente, o processo de montagem de *corpus* do *BootCaT* é composto de quatro passos:

1) construir um *corpus* automaticamente a partir de buscas no *Google*[26] utilizando um pequeno conjunto de itens léxicos, denominados sementes (*seeds*) no *BootCaT*;
2) extrair novas sementes desse *corpus*;
3) utilizar essas novas sementes para novas buscas ao *Google*, cujos textos recuperados serão concatenados ao *corpus*, aumentando-o;
4) extrair novas sementes desse *corpus* complementado-o, e assim por diante. A montagem de *corpus* proposta pelo *BootCaT* segue o diagrama da figura 8.

O primeiro passo é selecionar as sementes iniciais. Isso é feito manualmente, e boas sementes são termos típicos em textos do domínio específico do qual se busca construir a amostragem. No segundo passo, essas sementes são combinadas entre si e algumas dessas combinações (à escolha do usuário) são enviadas como buscas no *Google*. No terceiro passo, as URLs retornadas das buscas são processadas para obter-se apenas o texto contido nelas, convertendo-as para texto puro e "limpando-os", quando for possível. São aproveitados somente os formatos "html" e "txt". Nesse momento, um primeiro *corpus* já está formado. Desse primeiro *corpus* são extraídos *unigramas* (itens léxicos com apenas uma palavra), e a freqüência de cada unigrama obtido no *corpus* é apurada. Sabendo-se a freqüência de cada unigrama, esses podem ser comparados entre si. A relevância de cada unigrama é mensurada utilizando a medida estatística *log odds ratio* (Baroni e Bernardini, 2004), com o apoio de um *corpus* de referência na mesma língua. Uma lista de unigramas, ordenada pela relevância calculada pela medida *log odds ratio* é então gerada, e os primeiros elementos da lista são considerados bons candidatos a sementes. Caso o *corpus* obtido até o momento não seja satisfatório (seja pequeno, por exemplo), podem-se eleger os primeiros unigramas da lista como novas sementes e repetir o processo, voltando ao segundo passo. Segundo Baroni e Bernardini (2004), *corpus* representativos podem ser montados com poucas sementes iniciais (entre 5 e 15). Os autores também afirmam que com duas ou três iterações é possível obter um *corpus* satisfatório.

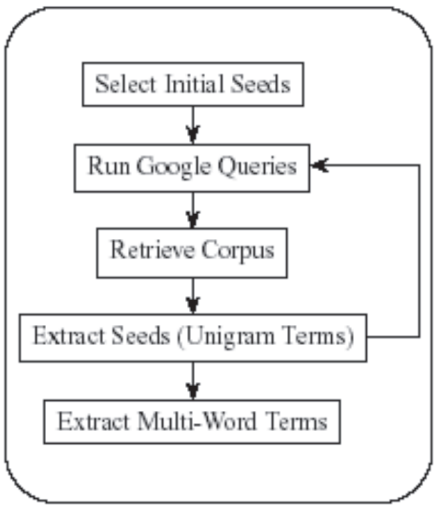


Figura 8. Fluxo de montagem de um *corpus* no *BootCaT* (Baroni e Bernardini, 2004).

O *BootCaT* também dispõe de ferramentas para extração de termos com mais de uma palavra, ou termos multipalavras. Para tal propósito, precisamos de duas listas, ambas obtidas no *corpus* de referência: uma de *conectores* e uma de *stopwords*. Conectores são compostos por palavras ou bigramas (itens léxicos com duas palavras, "meio ambiente", por exemplo) que ocorrem freqüentemente entre dois unigramas, e *stopwords* são termos muito freqüentes, geralmente formados por palavras de classe fechada de uma língua como os artigos, as conjunções, as preposições e os pronomes que não são conectores. As listas descritas acima não precisam necessariamente ser obtidas pelo *BootCaT*, podem ser dadas ou obtidas de outras fontes. Com as listas acima é possível definir o que são termos multipalavras, segundo as restrições abaixo:

1. contêm ao menos um unigrama;
2. não contêm *stopwords*;
3. podem ter conectores, desde que esses não estejam nas extremidades do termo e não sejam consecutivos;
4. têm freqüência maior que um limiar (*threshold*), que é relativo ao tamanho do termo;
5. não podem ser parte de termos multipalavras maiores com freqüência superior a *k*fq*, onde

[24] http://sslmit.unibo.it/~baroni/bootcat.html
[25] http://www.perl.com
[26] http://www.Google.com.br/

k é uma constante entre 0 e 1 (normalmente k é um valor perto de 1) e i *fq* é a freqüência do termo atual;

6. reciprocamente, não podem conter termos multipalavras menores com freqüência superior a (*1/k*) * *fq*;

Os termos multipalavras são procurados recursivamente, inicialmente buscando por bigramas e depois concatenando palavras à esquerda e à direita, na busca de um (n+1) grama. Parâmetros como a freqüência mínima para bigramas (utilizado para calcular o limiar da restrição 4) e o valor de *k* das restrições 5 e 6 devem ser informados pelo usuário.

O *BootCaT* é extremamente modular: para executar o processo de montagem de *corpus* e extração de termos são utilizadas várias ferramentas, sendo que o resultado de cada ferramenta serve de entrada para outra. Essa característica nos permite utilizar subconjuntos de ferramentas, conferir os arquivos de saída intermediários, adicionar novas ferramentas, substituir uma ferramenta ou alterar uma ferramenta sem preocupar-se com as outras, apenas cuidando para que ela aceite o mesmo tipo de entrada e produza o mesmo tipo de saída. Essa característica reduz re-implementações de algoritmos com implementações consolidadas, evitando a replicação desnecessária de código. Alterações intuitivamente complexas, como adaptações de ferramentas para trabalhar com línguas diferentes, têm sido experimentadas e comprovam os benefícios das ferramentas modulares. Adaptações para o *BootCaT* foram feitas para construção de *corpus* em língua japonesa (Baroni e Ueyama, 2004), com taxas encorajadoras de reaproveitamento de ferramentas e código.

As buscas e a recuperação das URLs dessas buscas requisitadas pelo *BootCaT* ao Google são possíveis por meio da API (Interface para Programação de Aplicativos) do Google. Essa API permite ao programador enviar e recuperar facilmente uma busca feita ao Google.

Para a utilização da API do *Google*, e conseqüentemente do *BootCaT*, é necessário obter a licença de uso dessa no *site* do *Google*. Para obter essa licença, o usuário precisa cadastrar-se, e a chave da licença é enviada por e-mail. Essa licença permite que o usuário execute diariamente até 1.000 buscas e retorne no máximo 10.000 resultados.

As ferramentas do BootCaT, por serem código livre, foram incorporadas no projeto e-Termos[27], uma aplicação *Computer-Supported Collaborative Work (CSCW)* composta por seis módulos de trabalho independentes, mas inter-relacionados, cujo propósito é automatizar ou semi-automatizar todas as tarefas de criação e gerenciamento do trabalho terminológico. O e-Termos, como um Ambiente Colaborativo é, *grosso modo*, um sistema Web cuja entrada principal é um *corpus* de especialidade de um determinado domínio do conhecimento; e a saída, um produto terminológico (glossário, dicionário, lista de termos, mapa conceitual, etc.) do domínio em questão. O e-Termos está sendo desenvolvido no NILC.

Lições aprendidas a partir de projetos de pesquisa

Vários projetos envolvendo *corpus* foram e têm sido objeto de pesquisa das autoras nos últimos anos. Apresentaremos, a seguir, detalhes da elaboração e execução desses projetos, com o intuito de expor detalhes da construção dos *corpora*, explicitando nossas escolhas, tomadas de decisão, erros cometidos, de forma a auxiliar demais pesquisadores que desejam adotar os princípios da Lingüística de *Corpus* em seus projetos.

Projetos “Corpus NILC” e “ Lácio-Web”

O NILC possui um *corpus* do português do Brasil (chamado de *Corpus* NILC ou CN), compilado a partir de 1993, contendo cerca de 35 milhões de palavras. O *corpus* consiste de textos em prosa, divididos em sub*corpora* de textos corrigidos, textos não corrigidos e textos semicorrigidos. As decisões de projeto e compilação foram motivadas pelas necessidades provenientes de outro projeto denominado ReGra[28] (um revisor gramatical para o português do Brasil, incorporado ao *Microsoft Word* desde 2000), embora na época as orientações da Lingüística de *Corpus* para compilação de *corpus* fossem incipientes. Alguns problemas do *Corpus* NILC são descritos abaixo (Pinheiro e Aluísio, 2003):

- classificação dos textos: a classificação textual do CN é problemática, pois o *Corpus* foi construído sob demanda. À medida que foram adquiridas, as amostras passaram a integrar categorias textuais distinguidas segundo parâmetros irregulares de classificação;
- quantidade de textos: alguns conjuntos de textos do CN são muito pouco representativos, isto é, não são quantitativamente suficientes em relação ao rótulo que carregam, como por exemplo: jornalístico, literário, jurídico, etc. O *corpus* científico, por exemplo, tem poucas amostras de teses, algumas dissertações incompletas e, de modo geral, é dedicado à área da informática. A quantidade de textos impede o aproveitamento do *corpus* para pesquisas gerais;

[27] O e-Termos está sendo desenvolvido por Leandro Henrique Mendonça de Oliveira, como tese de doutorado em Ciências de Computação e Matemática Computacional, com orientação de Sandra Maria Aluísio. O e-Termos foi originado do *TermEx*, projeto que será descrito a seguir (http://www.nilc.icmc.usp.br/etermos/).

[28] http://www.nilc.icmc.usp.br/nilc/projects/regra.htm

- compilação: alguns tipos de textos tiveram compilação irregular em relação ao padrão de amostragem aplicado em quase todo o CN. Embora o procedimento ideal de compilação fosse o de trazer apenas textos integrais, para algumas categorias essa regra foi quebrada, resultando em obras parcialmente compiladas;
- acúmulo de textos: uma característica insatisfatória de determinados conjuntos do CN é o acúmulo de textos em um único arquivo, resultado de uma escolha de formatação das amostras do *Corpus*. A opção foi a de anexar, num único arquivo, diversos textos pequenos, o que terminou ocultando especificidades sobre os textos, tais como as diferenças de autoria, de assunto, etc.

Para superar as limitações do CN, foi criado o Projeto Lácio-Web[29] (Aluísio *et al.*, 2003a, 2004).

O Lácio-Web (LW) foi um projeto financiado pelo CNPq, iniciado em 2002, com duração de 30 meses, e realizado em parceria entre o NILC, o Instituto de Matemática e Estatística (IME) e a Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH) ambos da USP, São Paulo. O objetivo do LW é divulgar e disponibilizar gratuitamente na Web: a) vários *corpora* do português brasileiro escrito contemporâneo, representando bancos de textos adequadamente compilados, catalogados e codificados, em um padrão que possibilita fácil intercâmbio, navegação e análise; e b) ferramentas lingüístico-computacionais, tais como contadores de freqüência, concordanciadores e etiquetadores morfossintáticos treinados em grandes *corpora* anotados manualmente.

O público-alvo do LW é heterogêneo: de um lado lingüistas, cientistas da computação, lexicógrafos, terminólogos, etc. e, de outro, o público em geral. O LW é acessado a partir de um portal (http://www.nilc.icmc.usp.br/lacioweb/), que informa os tipos de *corpus*, ferramentas, todo o material disponível e a forma de contribuir com textos para a continuação do projeto, disponibiliza, ainda, manuais e artigos relacionados e permite, após cadastramento do usuário, o acesso ao *corpus* e às ferramentas.

Dada a importância de um recurso de base como são os *corpora* de uma dada língua, para avançar estudos lingüísticos variados e também para a construção de sistemas computacionais de processamento de língua natural (PLN), justifica-se o sucesso que tivemos em conseguir permissão oficial para incluir materiais diversos, durante os 30 meses do projeto. Para obter essa permissão, foi incluído, juntamente com o termo de autorização, um texto explicativo apontando o potencial dos recursos e a necessidade de obtenção de textos integrais para diversas pesquisas lingüísticas, como por exemplo, a análise de textos e discursos e tarefas como a tradução.

O LW tenta preencher uma lacuna em termos de recursos para pesquisa e suporte à criação de ferramentas de PLN para a língua portuguesa do Brasil. Para tanto, quatro *corpora* foram disponibilizados: Lácio-Ref, Mac-Morpho, Par-C e Comp-C, descritos abaixo:

1) *Lácio-Ref*: *corpus* aberto e de referência composto de textos escritos em português brasileiro, respeitando a norma culta, com 4.278 arquivos, totalizando 8.291.818 ocorrências. É um *corpus* cru (não anotado com informações morfossintáticas, sintáticas ou de nível mais elevado), mas possui anotações da existência de elementos gráficos e anotação de cabeçalho. A grande maioria dos textos está disponibilizada na íntegra.

2) *Mac-Morpho*: *corpus* fechado e anotado morfossintaticamente, formado por artigos jornalísticos retirados da *Folha de S.Paulo*, ano 1994, dos cadernos Esporte (ES), Dinheiro (DI), Ciência (FC), Agronomia (AG), Informática (IF), Ilustrada (IL), Mais! (MA), Mundo (MU), Brasil (BR) e Cotidiano (CO). Composto de 1.167.183 ocorrências, o *corpus* foi etiquetado pelo analisador sintático *Palavras*, foi revisado manualmente quanto à anotação morfossintática e serviu de treinamento para três etiquetadores morfossintáticos disponíveis na Web (Aluísio *et al.*, 2003b). O MAC-MORPHO é disponibilizado para *download* em dois formatos: a) adequado para pesquisas lingüísticas com o uso de contadores de freqüência ou concordanciadores, por exemplo; b) adequado ao treinamento de etiquetadores e que, por ter as lexias complexas (multipalavras) separadas[30], teve o tamanho do *corpus* alterado para 1.221.468 ocorrências.

3) *Par-C*: *corpus* aberto, paralelo, Inglês-Português, que possui, inicialmente, textos de um ano de edições da revista *Pesquisa Fapesp*, num total de 646 textos em cada língua. O número total de ocorrências desse *corpus* é de 893.283.

4) *Comp-C*: *corpus* aberto, formado por textos originais de conteúdo comparável em inglês e português, inicialmente disponível apenas para o gênero jurídico. Conta com 29 textos, 61.149 ocorrências, e será ampliado futuramente. Os *corpora* comparáveis são projetados para a avaliação de métodos de extração de termos para sistemas de PLN, para confecção de glos-

[29] Coordenado por Sandra Maria Aluísio (ICMC/USP).

[30] "Rio=de=Janeiro_NPROP", por exemplo, é separado em "Rio_NPROP de_NPROP Janeiro_NPROP", em que NPROP é uma etiqueta para nomes próprios.

sários e dicionários especializados e para outras pesquisas lingüísticas.

No total, o Projeto LW possui 5.708 arquivos, totalizando 10.413.524 ocorrências.

O LW distingue seus textos em quatro categorias ortogonais: gênero, tipo de texto, domínio e meio de distribuição. A definição e a composição das categorias são detalhadas abaixo.

- *Gênero textual*: para o Projeto Lácio-web, o gênero discrimina o texto pela intenção comunicativa e pelo caráter discursivo, isto é, a comunidade (meio) em que circula e as atividades humanas é que o tornam relevante. Convencionamos o uso de um super-gênero, chamado Literário (LT), um conjunto de gêneros e um conjunto de subgêneros. Os gêneros e subgêneros são dados no Quadro 1.

Quadro 1. Gêneros e subgêneros utilizados no Projeto Lácio-web.

Gênero	Subgênero
Científico (CI)	———
De referência (RE)	enciclopédico, lexicográfico, terminológico e outros.
Informativo (IF)	jornalístico e outros
Jurídico (JU)	———
Prosa (PR)*	biografia, conto, novela, romance e outros
Poesia (PO)*	———
Drama (DR)*	———
Instrucional (IS)	didático, procedimental e outros
Técnico-Administrativo (TA)	———

* Esses gêneros, especialmente, advêm do supergênero Literário.

- *Tipo textual*: considera-se "tipo de texto" o modo específico de estruturação de um texto. Refere-se ao texto visto "de dentro", ou seja, suas partes componentes, seu léxico, sua sintaxe, sua adequação ao tema etc. Trata-se de uma lista em constante atualização e que, no momento, é composta de 39 categorias (e "Outros" – tipos textuais não previstos), por ex.: apostila, manual, parecer, reportagem, súmula, testamento etc.
- *Domínio*: é a "área de conhecimento" que tematiza a principal informação veiculada pelo texto. Temos três grandes linhas de domínio, denominadas "domínio geral". A cada uma dessas linhas associam-se subdomínios, denominados "domínios específicos". A divisão em termos de domínio geral apresenta as seguintes subdivisões:
  - *a) científica*: refere-se aos textos de ciências. Esse grupo é composto por seis áreas do conhecimento: Ciências Agrárias, Ciências Biológicas, Ciências da Saúde, Ciências Exatas e da Terra, Ciências Humanas e Ciências Sociais Aplicadas;
  - *b) religião e pensamento*: envolve os temas metafísicos, espirituais e teológicos (ex.: livros de bruxaria, de auto-ajuda, etc.).
  - *c) generalidades*: refere-se aos textos com temas variados e, de modo geral, inseridos num campo conhecido pelo senso comum (ex.: entretenimento). Inclui, além disso, os textos que abordam, de forma não-analítica, temas considerados pela ciência (exs.: ciência e tecnologia, saúde, esporte, etc.).
- *Meio de distribuição*: seleciona o canal por meio do qual o texto foi divulgado ao seu público-alvo, por ex.: CD_ROM (CR), Diário Oficial (DO), Internet (IN), Jornal (JO), Livro (LI), Tese (TE).

O Projeto Lácio-Web disponibiliza várias ferramentas lingüístico-computacionais como concordanciadores, contadores de freqüência e etiquetadores morfossintáticos, treinados com o *corpus* do português do Brasil e anotado manualmente – o MAC-Morpho. O objetivo é facilitar a observação de características lingüísticas do português do Brasil por pesquisadores, assim como melhorar a qualidade dos sistemas desenvolvidos para o português, tais como, tradutores, sumarizadores automáticos e máquinas de busca (como o *Google*, por exemplo).

As ferramentas podem ser usadas com o Lácio-Ref, com os sub*corpora* criados pelo usuário ou ainda com o *corpus* que o usuário tiver carregado para o LW, pois é possível fazer *upload* de textos.

Ao final dos 30 meses de pesquisa e desenvolvimento, o LW disponibiliza, de forma gratuita, amigável e de largo alcance: a) quatro tipos distintos de *corpus* (Lácio-Ref, Mac-Morpho, Par-C e Lácio-Dev); b) algumas ferramentas de processamento lingüístico-computacional (contador de freqüência, concordanciador e etiquetador morfossintático); e c) um Portal que, sensível a diferentes tipos de usuários, oferece três tipos de interface de pesquisa, com ferramentas de base associadas, sendo, além disso, um ambiente de navegação dinâmica, didática e, sobretudo, de incentivo ao uso de *corpus* para os mais diversos tipos de investigação lingüística, uma vez que permite o *download* completo das amostras dos *corpora*. Mas ainda assim o LW deixou lacunas importantes como a falta de um balanceamento de *corpus*, como, por exemplo, em gênero e número de textos por categorias. Várias decisões tomadas no projeto LW ainda estão um pouco distantes dos padrões internacionais, como o XCES (Ide *et al.*, 2000), tanto com relação à anotação como à

codificação, embora tenhamos dado um grande passo em direção à padronização com a proposta de um rico cabeçalho em XML que traz informações bibliográficas e da tipologia quadripartida; e a anotação explícita da existência de elementos gráficos retirados dos textos.

Projeto TermEx

O projeto[31] intitulado *Extração automática de termos e elaboração colaborativa de terminologias para intercâmbio e difusão de conhecimento especializado (TermEx)* foi financiado pela FAPESP, iniciou-se em 2003 e encerrou-se em 2005. O projeto foi uma parceria entre a UFSCar e a USP/São Carlos e tinha como principais objetivos: 1) pesquisar e implementar métodos para a extração automática de termos; 2) criar um ambiente computacional para auxílio na pesquisa terminológica/terminográfica; 3) elaborar um dicionário terminológico para a área de revestimento cerâmico.

Como nossa proposta final era a elaboração de um dicionário terminológico, o *corpus* foi elaborado a partir de artigos especializados da revista *Cerâmica Industrial*[32]. Essa revista, escrita em português, tem como objetivo contribuir para atualização e melhoria da formação dos técnicos cerâmicos brasileiros. É destinada fundamentalmente a profissionais da indústria. Os especialistas que colaboram com artigos são tanto pesquisadores (brasileiros e estrangeiros) de laboratórios, institutos de pesquisas e desenvolvimento (P&D) e universidades, quanto profissionais que atuam em indústrias. Constitui uma publicação bastante relevante e respeitada no setor de Revestimento Cerâmico. Daí a nossa escolha, já que uma das nossas preocupações era abarcar não só a linguagem utilizada nos laboratórios e institutos de P&D como também aquela utilizada nas indústrias. Acreditávamos que a escolha dessa revista satisfazia os requisitos *representatividade* e *amostragem*.

Os textos foram agrupados pelos anos em que foram publicados, 1996-2003, e totalizam 196, possuindo, cada texto, uma média de sete a oito páginas (aproximadamente 4.000 palavras). Todos os textos presentes no *site* da revista estão no formato "pdf". Porém, para que eles pudessem ser processados pelos métodos propostos nesse trabalho, deveriam estar no formato "txt". Por essa razão, nem todos os textos foram utilizados, visto que ocorreram alguns problemas no processo de conversão do formato "pdf" para "txt", o que totalizou 164 textos.

Percebemos, entretanto, que embora todos fossem escritos em português, 55 desses artigos eram de autores estrangeiros, quatro escritos por autores estrangeiros e nacionais, e quatro cuja nacionalidade era desconhecida. Diante dessas constatações, a montagem do *corpus* foi reavaliada, pois isso afetaria o requisito *autenticidade*. A retirada desses textos, por outro lado, comprometeria a *extensão* do *corpus*, uma vez que uma das abordagens de extração de termos que seria utilizada era a estatística, abordagem dependente, significativamente, do tamanho do *corpus*. Contatamos, então, o responsável pela revista para esclarecer se esses textos, depois de traduzidos, eram revisados por um especialista falante nativo do português. Como a resposta foi afirmativa, todos aqueles textos, objeto de preocupação, foram incluídos no *corpus*. Observe-se que, neste caso, demos prioridade para o requisito *extensão* em detrimento da *autenticidade*.

Para a transformação dos textos para o formato TXT, foi utilizada a ferramenta denominada EXTEX (Extracção de Texto de Ficheiros Formatados)[33]. Uma característica dessa ferramenta, ao realizar a transformação, é a de que o texto transformado não é totalmente igual ao texto original. Ele se apresenta com junção de algumas palavras, preserva os índices de referência bibliográfica e as notas de rodapé anexadas às palavras, e a hifenização dos textos no formato "pdf". Para resolver esses problemas, esses textos foram submetidos a um processo cuidadoso de correção manual.

Vale ressaltar também que todos os arquivos do *corpus* foram pré-processados para a retirada de informações de autoria e filiação, referências bibliográficas, figuras, tabelas e quadros, fazendo com que o tamanho médio dos artigos diminuísse de oito para cinco páginas, totalizando 448.352 palavras.

Também foi encontrada grande quantidade de erros gramaticais e de digitação. Para minimizar os erros gramaticais, foi realizada uma varredura no *corpus* com o auxílio de um processador de textos, buscando corrigir os erros encontrados, podendo-se, dessa forma, analisar os dados de forma mais precisa.

O *corpus* foi pré-processado utilizando-se um *tokenizador*[34] desenvolvido no NILC[35] chamado *Sentencer*, que é um *tokenizador* e segmentador sentencial para português, que *tokeniza* um texto de entrada, inserindo um caractere de fim de linha ao fim de cada sentença. Linhas em branco marcam fronteiras de parágrafo. Apenas caracteres de fim de linha, como ponto-final, ponto-de-interrogação, ponto-de-exclamação e reticências são considerados possíveis finais de sentença. O programa *Sentencer* trata de abreviações como "Dr.", "Prof.", não considerando, nesse caso, o ponto final como um caractere de fim de linha, ao contrário, o ponto é desconsiderado. Além disso, o programa *Sentencer* também apresenta a

[31] O projeto foi coordenado por Gladis Maria de Barcellos Almeida (UFSCar) e contou com a colaboração de Sandra Maria Aluísio (USP).
[32] http://www.ceramicaindustrial.org.br/.
[33] http://poloclup.linguateca.pt/ferramentas/extex/
[34] Ferramenta computacional que separa o texto em *tokens* (palavra, ponto, espaço, qualquer sinal gráfico).
[35] http://www.nilc.icmc.usp.br/nilc/

função de separar os caracteres (como aspas, vírgulas, pontuações, entre outros) dos *tokens*.

Após o *corpus* ter sido *tokenizado* pelo *Sentencer*, ele foi etiquetado[36] utilizando-se o MXPOST (Ratnaparkhi, 1996), etiquetador que foi treinado no NILC com um conjunto simplificado que possui 15 etiquetas[37] e um *corpus* manualmente etiquetado de 104.963 palavras. Esse etiquetador[38] possui a precisão de 97%. Para usar o MXPOST no arquivo de entrada, cada *token* deveria estar separado por um espaço em branco, ou seja, nenhum caractere, incluindo pontuação, deveria estar anexo às palavras; essa foi uma das razões para o uso do programa *Sentencer*.

Após o pré-processamento, o *corpus* estava pronto para ser objeto de extração automática de termos.

Antes de realizar a extração, alguns métodos automáticos foram avaliados e implementados para o português[39], especificamente métodos das três abordagens para o português: estatística, lingüística e híbrida.

Os métodos baseados em conhecimento estatístico geralmente detectam as unidades terminológicas de acordo com a freqüência com que elas ocorrem em um *corpus*. Existem métodos estatísticos que utilizam desde simples freqüências até aqueles que utilizam estatísticas mais complexas, como informação mútua e coeficiente *log-likelihood* e *c-value*. A função é, em todos os métodos, identificar os candidatos a termo (Teline *et al*., 2003).

Os sistemas baseados em conhecimento lingüístico utilizam diferentes recursos que contêm diferentes informações lingüísticas para a extração dos termos. Essas informações lingüísticas dizem respeito a: informações lexicográficas – dicionários de termos e lista de palavras auxiliares ("*stopwords*"); informações morfológicas – padrões de estrutura interna da palavra; informações morfossintáticas – categorias morfossintáticas e funções sintáticas; informações semânticas – classificações semânticas; informações pragmáticas – representações tipográficas e informações de disposição do termo no texto. Este tipo de conhecimento utilizado faz com que os sistemas baseados em conhecimento lingüístico se apliquem somente a uma língua e, às vezes, até mesmo a uma única variante (Teline *et al*., 2003).

Os sistemas baseados em conhecimento híbrido utilizam o conhecimento estatístico juntamente com o lingüístico. A aplicação do conhecimento híbrido torna o sistema mais eficiente, visto que ele condiciona os resultados. Existem dois tipos de métodos híbridos: aqueles que aplicam o conhecimento estatístico primeiro e depois o lingüístico, e aqueles que utilizam a estatística apenas como um complemento da lingüística (Teline *et al*., 2003).

Como o trabalho de Teline (2004) atestou que os sistemas baseados em conhecimento híbrido eram os mais eficientes, optou-se por essa abordagem no projeto TermEx. Ocorre que o léxico[40] utilizado para o reconhecimento das estruturas morfolexicais da terminologia de Revestimento Cerâmico era constituído de itens da língua geral, o que acabou impedindo que esse léxico reconhecesse determinados termos. Observe-se como o léxico do *ReGra* lematizou determinados termos multipalavras: *ação mecânica* > *ação mecânico, alumina calcinada* > *alumina calcinar, capacidade instalada* > *capacidade instalar*. Em vista desse cenário, utilizamos então a abordagem estatística.

Uma grande lição que aprendemos com o projeto *TermEx* foi o fato de não termos balanceado o *corpus* de forma a incluir distintos gêneros. Esse erro foi observado posteriormente quando procurávamos contextos definitórios ou explicativos para elaborarmos as definições para o dicionário. Nossa hipótese era de que um *corpus* contendo apenas textos do gênero técnico-científico fosse suficiente para a elaboração de um dicionário terminológico. Entretanto, quando os autores escrevem um artigo científico, têm como público-alvo leitores especialistas que não necessitam de explicações conceituais de objetos, maquinário, conceitos, técnicas, etc. As glosas, portanto, estão ausentes desse tipo de texto. Vamos encontrar contextos definitórios ou explicativos nos gêneros científico de divulgação e instrucional (apostila, livro-texto, manual, por exemplo). A constatação a que chegamos é que mesmo em se tratando de uma pesquisa terminológica, o *corpus* deve ser balanceado, contendo, pelo menos, textos desses três gênero: técnico-científico, científico de divulgação e instrucional. Percebemos que a falta de balanceamento acabou gerando um *corpus* menos representativo, com menos amostras e menos diversificado, erros que não devem ser repetidos, posto que esse *corpus* afetou diretamente a redação dos verbetes.

---

[36] Etiquetar significa classificar o texto morfologicamente, ou seja, atribuir a cada unidade a classe correspondente.

[37] I-interjeição; LOCU-locução; PREP-preposição; N-substantivo; NP-nome próprio; VERB-verbo; ADJ-adjetivo; AUX-verbo auxiliar; ADV-advérbio; PRON-pronome; CONJ-conjunção; NUME-numeral; ART-artigo; RES- resíduo; PDEN-palavra denotativa e mais 4 tipos de contrações: PREP+ART, para palavras como "da", "na"; PREP+PD, para palavras como "nesta", "naquela", "nessa"; PREP+PPR, para palavras como "dela", "nela"; PREP+N, para palavras como "d'alma", "d'água", "d'arte".

[38] O NILC dispõe de vários etiquetadores que podem ser acessados a partir de http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html.

[39] A avaliação e a implantação dos métodos foi objeto de um trabalho de mestrado, denominado *Avaliação de métodos para extração automática de terminologia de textos em português (ExPorTer)* (Teline, 2004).

[40] O léxico utilizado foi o do ReGra (Revisor Gramatical do Português), que contém 68.530 lemas e 1.563.136 entradas, incluindo formas flexionadas, palavras compostas e locuções (5.763 das entradas são locuções nominais, prepositivas, adjetivas, adverbiais, conjuntivas). Esse léxico está em constante atualização (desde 1993), já que é o léxico que dá suporte ao corretor sintático do *Microsoft Word*. O ReGra foi desenvolvido no NILC (www.nilc.icmc.usp.br/nilc/projects/regra.htm).

Projeto NanoTerm

O projeto[41] intitulado *Terminologia em Língua Portuguesa da Nanociência e Nanotecnologia: Sistematização do Repertório Vocabular e Elaboração de Dicionário-Piloto* (*NanoTerm)* é financiado pelo CNPq e foi iniciado em 2006 (com vigência de dois anos). O projeto é também uma parceria entre a UFSCar e a USP/São Carlos e tem como objetivos: 1) a constituição de um *corpus* em língua portuguesa da Nanociência e Nanotecnologia (N&N); 2) a busca de equivalentes em português (língua de chegada) a partir de uma nomenclatura em inglês (língua de partida); 3) uma ontologia em língua portuguesa da área de N&N; 4) a elaboração do primeiro dicionário-piloto de N&N em língua materna.

Para a construção do *corpus*, inicialmente, foi realizado um estudo exploratório dos textos existentes em língua portuguesa bem como dos gêneros aos quais eles pertencem. Embora tivéssemos tentado balancear o *corpus*, inserindo uma quantidade equilibrada de textos dos gêneros informativo, científico de divulgação e científico, obtivemos uma grande quantidade dos primeiros e uma quantidade reduzida do último (científico). Entendemos que isso se deve fato de a área de N&N ser relativamente nova no Brasil, além disso, os pesquisadores, fundamentalmente das áreas de Exatas e Biomédicas que atuam em N&N, publicam seus resultados de pesquisa em língua inglesa. Os tipos de textos que compõem o gênero CIENTÍFICO são fundamentalmente dissertações e teses.

Ressalte-se que até o momento todos os textos foram obtidos na Web. É importante destacar que muitas páginas da *Internet*, embora se tivessem revelado útil para a pesquisa, estavam acessíveis somente para sócios ou assinantes, inviabilizando, portanto, a obtenção dos textos. Serão ainda inseridos no *corpus* textos impressos, os quais serão posteriormente digitalizados. No estudo exploratório que fizemos, encontramos apenas dois livros, cinco artigos e um relatório. Evidentemente, será necessário insistir na busca por mais textos impressos.

Após a seleção dos textos, foi realizada a compilação dos textos obtidos na Web. Para essa compilação, foram utilizados os seguintes itens de busca: *nanociência*, *nanotecnologia*, *genômica.* Todavia, após realizarmos buscas, decidimos incluir e o prefixo *nano-* para abarcar termos como: *nanotubo, nanorrede/nano-rede*, *nanocápsula, nanoesfera, nanobiotecnologia*, etc. Assim que cada texto era compilado, procedia-se com a sua manipulação, isto é, com a conversão manual e automática (Pacote XPDF[42]) de formatos "doc", "html" e "pdf" para "txt" e na limpeza e formatação.

Depois que todos os textos foram convertidos em formato "txt", eles receberam uma nomeação, de acordo com um padrão previamente determinado, de forma a facilitar a recuperação posterior de cada texto. Após a nomeação dos arquivos, foi gerado (de forma semi-automática) um cabeçalho para cada texto. A geração semi-automática desse cabeçalho foi feita por meio de um *editor* (programa computacional "com interface gráfica" para criar ou modificar arquivos) que auxilia o lingüista a especificar diversas informações sobre os textos. Ressaltamos que esse programa é uma versão adaptada no Editor de Cabeçalho utilizado no Projeto Lácio-Web[43] e contém os seguintes campos de informação: título, subtítulo, fonte, editor, local de publicação, data, assunto, autoria, tipo de autoria (individual ou coletiva), sexo do autor, tipo de texto, meio de distribuição e comentários (introduzem-se nesse campo informações adicionais sobre o texto). Observe-se, nas Figuras 9 e 10, algumas telas do editor de cabeçalho que pode ser obtido gratuitamente na página do projeto Lácio-Web.

Para cada texto, é gerado um cabeçalho. É possível ver na Figura 11 como ficam as informações anotadas em XML. São essas informações anotadas em XML que vão permitir posteriormente que se façam buscas específicas.

O preenchimento de todos esses campos do cabeçalho é útil para esta pesquisa porque a partir desses dados será possível fazer constatações tais como: o repertório vocabular tem alguma relação com a temática do texto, com o gênero, com a autoria ou com o meio de distribuição? Dependendo do tema tratado em determinado texto, é possível recuperar os descritores desse texto por meio da freqüência? Em outras palavras: num texto cujo tema seja Nanociência, o item léxico *nanociência* ocorre quantas vezes? Enfim, além das buscas que poderão ser empreendidas por cada campo constitutivo do cabeçalho, é possível fazer constatações relevantes sobre o léxico.

Ao final de processo de construção do *corpus*, o projeto *NanoTerm* deverá totalizar cerca de um milhão de palavras.

Projeto Dicionário Histórico ~~- nós~~

O projeto[44] intitulado *Dicionário Histórico do Português do Brasil (séculos XVI, XVII e XVIII)*, no âmbito do programa "Institutos do Milênio" do CNPq, é financiado por este órgão e iniciou-se em dezembro de 2005 (com vigência de 3 anos). A equipe envolvida no projeto conta com 10 universidades, 17 doutores e 17 alunos de graduação e pós-graduação. O projeto tem como principal objetivo a elaboração de um dicionário do português cor-

[41] O projeto é coordenado por Gladis Maria de Barcellos Almeida (UFSCar) e conta com a colaboração de Sandra Maria Aluísio (USP).
[42] XPDF é um programa de código aberto que permite a conversão automática de arquivos, conferir: http://www.foolabs.com/xpdf/.
[43] http://www.nilc.icmc.usp.br/lacioweb/
[44] O projeto é coordenado por Maria Tereza Camargo Biderman (UNESP/campus de Araraquara).

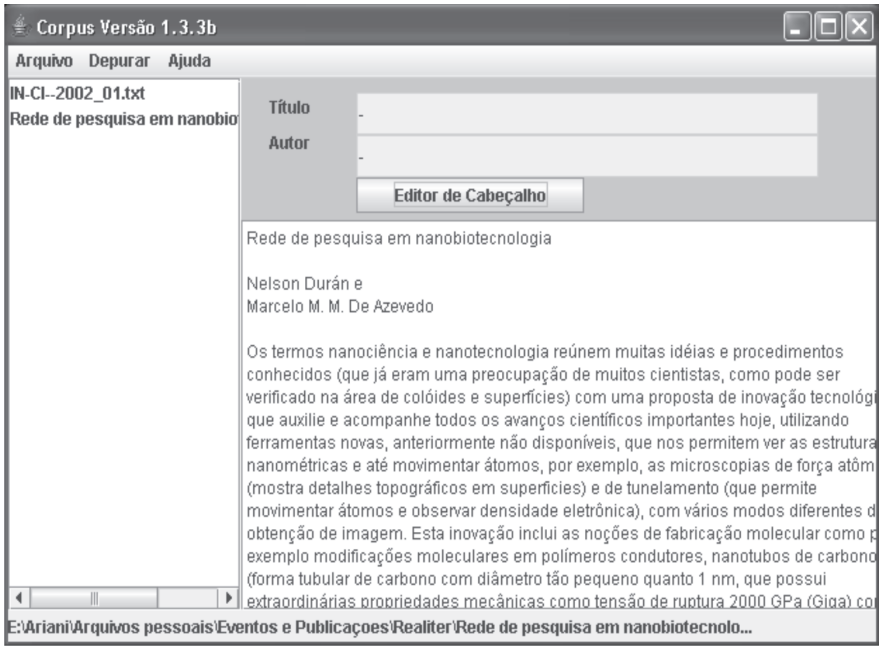


Figura 9. Editor de cabeçalho adaptado do projeto Lácio-Web.

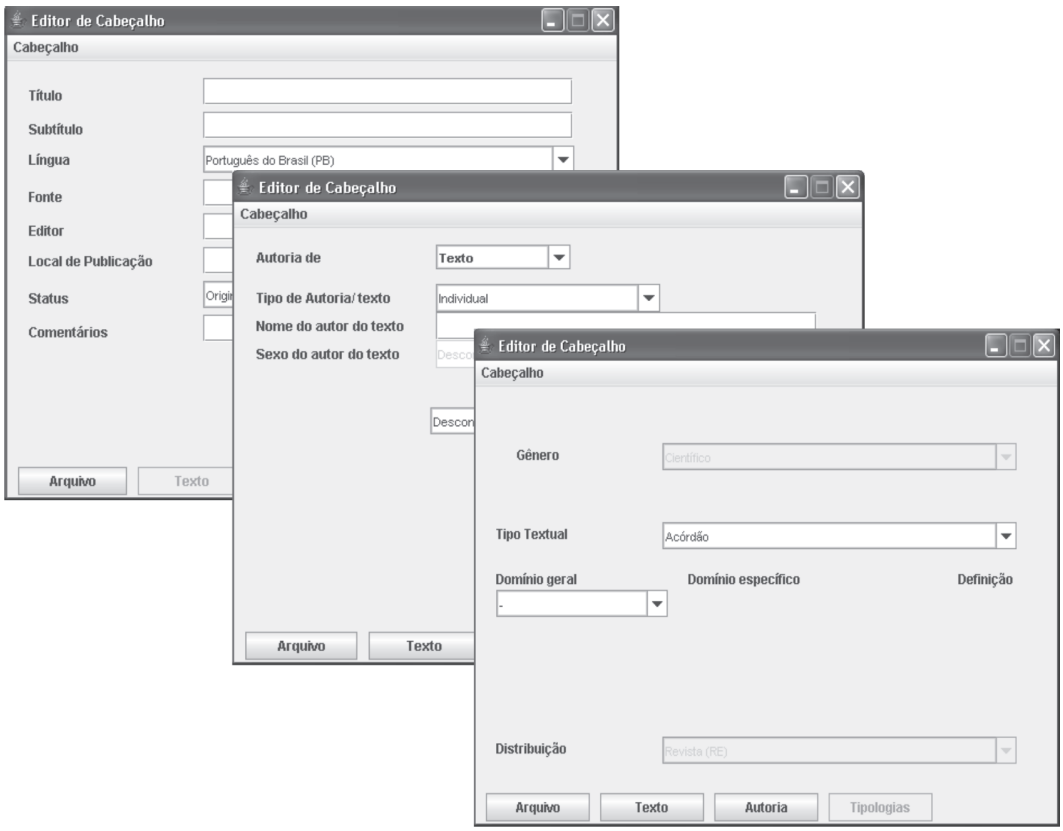


Figura 10. Janelas do editor para a especificação de informações bibliográficas, de autoria e da tipologia quadripartida (gênero, tipo textual, domínio e meio de distribuição).

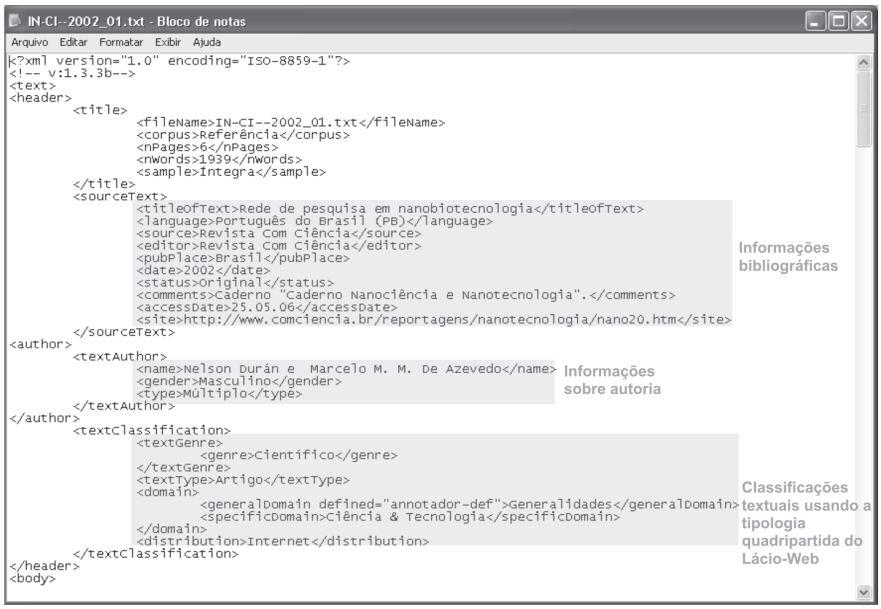


Figura 11. Cabeçalho com etiquetas XML gerado pelo Editor de Cabeçalho do Lácio-Web.

respondente aos séculos XVI, XVII e XVIII. Convém assinalar que o Brasil não conta com nenhuma obra lexicográfica sobre seu vocabulário nos primeiros tempos da formação do Português Brasileiro, o que atesta a originalidade da proposta.

Para a elaboração desse dicionário, é necessária a construção de um *corpus*, evidentemente.

De forma a cumprir os seis requisitos citados no início deste artigo (autenticidade, representatividade, balanceamento, amostragem, diversidade e extensão), o *corpus* está sendo construído obedecendo a uma seqüência de etapas.

Foi realizada inicialmente a seleção dos textos. Essa seleção tem como orientação os seguintes pontos: a) os textos têm de ser escritos originalmente em português por indivíduos nascidos no Brasil, se tiverem nascido em Portugal, teriam de estar residindo no Brasil há anos (autenticidade), embora saibamos que há pouco material disponível com essas características no século XVI; b) seleção de documentos de forma a abarcar distintos domínios do saber, gêneros discursivos e tipologias textuais (representatividade, balanceamento, amostragem, diversidade); c) distribuição desses gêneros e domínios nos três séculos que envolvem a pesquisa, por exemplo, o gênero literário só será pertinente no século XVIII, posto que antes disso não se pode afirmar que havia uma literatura genuinamente brasileira (balanceamento); d) seleção de uma quantidade de textos suficientes para a elaboração de um dicionário que contemple a diversidade lexical desses séculos (extensão), no que se refere às classes abertas, a saber: substantivo, adjetivo, verbo e advérbio. A previsão inicial é de que o *corpus* conte-

nha, no mínimo, 3 milhões de palavras, para gerar, pelo menos, dez mil entradas no dicionário.

A construção desse *corpus* inicia-se com o processo de digitalização, já que os textos referentes a esses séculos estão, em sua grande maioria, na forma impressa.

Após a análise e seleção das obras, os livros são digitalizados em formato de imagem (arquivos de imagem com extensão "tiff") para, então, serem transformados em textos (arquivos de texto com extensão "doc"). Depois que estão em formato "doc", os textos passam por um processo de revisão manual. Este é um trabalho minucioso e que requer muita atenção, pois se trabalha com a leitura cotejada de 3 documentos: a) a imagem do texto original, em forma de figura (extensão "tiff") gerada por digitalização; b) a imagem do texto digitalizado em forma de texto propriamente (em formato "doc"); c) o texto original impresso que deve estar sobre a mesa, à mão, para o caso de a imagem no computador não ser suficiente para dirimir dúvidas. Se os textos fossem atuais, a tarefa estaria terminada, contudo, é importante lembrar que estamos trabalhando com textos antigos e que a dificuldade está justamente na grafia não padronizada do português quinhentista.

É importante assinalar que a digitalização exige alguns cuidados, pois os documentos possuem normalmente páginas em papel pardo, muito amarelas ou com manchas próprias do envelhecimento, folhas craqueladas, páginas soltas, etc. Toda essa "sujeira" na imagem pode implicar a geração de caracteres estranhos ou falhas no texto digitalizado que precisam ser eliminadas durante a revisão. Assim, após a digitalização, é preciso limpar e recortar cada uma das imagens digitalizadas para que elas as-

sumam um formato padrão o mais "limpo" possível, isso tornará a fase de revisão manual menos penosa.

Todo o material digitalizado é organizado de forma que cada unidade de texto constitua dois arquivos: um em forma de imagem e o seu correspondente em forma de texto. Cada texto possui um extenso cabeçalho e é organizado em pastas que correspondem à determinada obra. Por exemplo, a obra *Tratado Descritivo do Brasil*, de Gabriel Soares Sousa, após a digitalização, foi transformada em 24 arquivos "tiff" e, depois da revisão, passou a ter também 24 arquivos "doc". Isso significa que após um ano de trabalho o projeto contará com um *corpus* e com um banco de imagens "tiff" correspondendo a cada texto.

É a partir do formato "doc" que os textos estão prontos para receberem outros tratamentos possibilitando o processamento computacional. Como os textos possuem caracteres que não pertencem ao conjunto ANSI[45], é necessário a sua codificação utilizando o *Unicode*, que uniformiza vários conjuntos de caracteres para muitas línguas, inclusive as línguas orientais.

Considerações finais

Neste artigo, procuramos apresentar a concepção de *corpus* para a Lingüística e para a Lingüística de *Corpus*, abordar questões importantes para a elaboração de *corpus* computadorizado, discorrer sobre as etapas metodológicas para a compilação de *corpus*, citar alguns *corpora* e ferramentas disponíveis na Web para pesquisa e construção de *corpus*, e, finalmente, detalhar quatro projetos de pesquisa envolvendo *corpus*, de forma a auxiliar demais pesquisadores que desejam adotar os princípios da Lingüística de *Corpus* em seus projetos.

Nosso intuito foi oferecer um panorama das práticas da Lingüística de *Corpus*. Esperamos que essas reflexões e relatos possam nortear as pesquisas, levantar mais questionamentos e sedimentar as práticas da Lingüística de *Corpus* no Brasil.

Referências

ALUÍSIO, S.M.; PINHEIRO, G.; FINGER, M.; NUNES, M.G.V. e TAGNIN, S.E.O. 2003a. The Lácio-Web Project: overview and issues in Brazilian Portuguese *corpus* creation. *In*: *CORPUS* LINGUISTICS 2003, Lancaster, UK, 2003. *Proceedings*...Lancaster, UCREL - Lancaster University, **16**:14-21. (Also as UCREL Technical Report, Vol 16 Part).

ALUÍSIO, S. M.; PELIZZONI, J. M.; MARCHI, A. R.; OLIVEIRA, L. H.; MANENTI, R. e MARQUIVAFÁVEL, V. 2003b. An account of the challenge of tagging a reference *corpus* of Brazilian Portuguese. *In*: PROPOR´2003, Faro, Portugal, 2003. *Proceedings*... Lecture Notes in Computer Science. New York, Springer, **1**:110-117.

ALUÍSIO, S.M.; PINHEIRO, G.M.; MANFRIM, A.M.P.; OLIVEIRA, L.H.M. de; GENOVES Jr., L.C. e TAGNIN, S.E.O. 2004. The Lácio-Web: *Corpora* and Tools to Advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. *In*: LREC 2004, Lisboa, Portugal, 2004. *Proceedings*... Paris, ELDA, p. 1779-1782.

ATKINS, S.; CLEAR, J. e OSTLER, N. 1992. Corpus design criteria. *Journal of Literary and Linguistic Computing*, **7**(1).

BARONI, M. e BERNARDINI, S. 2004. BootCaT: Bootstrapping corpora and terms from the web. In: LREC 2004, Lisboa, Portugal, 2004. *Proceedings*...Paris, ELDA.

BARONI, M. e UEYAMA, M. 2004. Retrieving Japanese specialized terms and corpora from the World Wide Web. *In*: KONVENS, Viena, Áustria, 2004. *Proceedings*...Viena, OFAI.

BERBER SARDINHA, T. 2000. Histórico e problemática. *D.E.L.T.A.*, **16**(2):323-367.

BERBER SARDINHA, T. 2004. *Lingüística de corpus*. São Paulo, Manole, 410 p.

BIBER, D. 1993. Representativeness in Corpus Design**.** *Lit Linguist Computing*, **8**:243-257.

BIBER, D.; CONRAD, S. e REPPEN, R. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press, Cambridge.

DUBOIS, J; GIACOMO, M.; GUESPIN, L.; MARCELLESI, C.; MARCELLESI, J.B. e MEVEL, J.P. 1993. *Dicionário de lingüística*. São Paulo, Cultrix, 653 p.

DUCROT, O. e TODOROV, T. 2001. *Dicionário enciclopédico das ciências da linguagem*. 3ª ed., São Paulo, Perspectiva, 339 p.

GALISSON, R. e COSTE, D. 1983. *Dicionário de didáctica das línguas*. Coimbra, Livraria Almedina, 763 p.

HASUND, K. 1998. Protecting the innocent: The issue of informants' anonymity in the COLT corpus. *In*: A. RENOUF (ed.), *Explorations in Corpus Linguistics*. Amsterdam, Rodopi, p. 13-27.

IDE, N.; BONHOMME, P. e ROMARY, L. 2000. XCES: An XML-based Standard for Linguistic *Corpora*. *In*: Second Language Resources and Evaluation Conference (LREC), Athens, Greece, 2000. *Proceedings*..., p. 825-830.

KENNEDY, G. 1998. *An Introduction to Corpus Linguistics***.** London;New York, Longman.

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.

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. 2002. *Manuel d'utilisation du logiciel Unitex*. IGM, Universite de Marne-la-Vallee, 217 p. Disponível em: http://www-igm.univ-mlv.fr/~unitex/. Acesso em: 20/10/2006.

PINHEIRO, G.M.e ALUÍSIO, S.M. 2003. Córpus Nilc: descrição e análise crítica com vistas ao projeto Lacio-Web. *NILC-TR-03-03*, fevereiro, 60 p.

RATNAPARKHI, A. 1996. A Maximum Entropy Part-Of-Speech Tagger. *In*: Empirical Methods in Natural Language Processing Conference, Philadelphia, Pennsylvania,1996. *Proceedings*... Philadelphia, University of Pennsylvania, p. 133-142.

RENOUF, A. (ed.). 1998. *Explorations in Corpus Linguistics*. Amsterdam, Rodopi.

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.

TELINE, M.F. 2004. *Avaliação de métodos para extração auto-*

[45] American National Standards Institute – ANSI (http://www.ansi.org/)

mática de terminologia de textos em português. São Carlos, SP. Dissertação de mestrado. Universidade de São Paulo – USP, 136 p.

TELINE, M.F.; ALMEIDA, G.M.B. e ALUÍSIO, S.M. 2003. Extração manual e automática de terminologia: comparando abordagens e critérios. *In*: Workshop em Tecnologia da Informação e da Linguagem Humana, **1**, São Carlos, SP, 2003. *Anais*... São Carlos, USP. (CD-ROM).

TRASK, R.L. 2004. *Dicionário de Linguagem e Lingüística.* São Paulo, Contexto, 364 p.

VALE, O.A. 1998. Sintaxe, léxico e expressões idiomáticas. *In*: A.N. BRITO e O.A. VALE (orgs.), *Filosofia, lingüística, informática: aspectos da linguagem*. Goiânia, Editora UFG, p. 127-137.

VALE, O.A. 2001. *Expressões Cristalizadas do Português do Brasil: uma proposta de tipologia*. Araraquara, SP. Tese de doutorado. Universidade Estadual Paulista – UNESP.

Submetido em: 10/2006
Aceito em: 11/2006

Sandra Maria Aluísio

Doutora em Física e Pós-Doutorado em Ciências da Computação. Professora efetiva da USP, Brasil

Gladis Maria de Barcellos Almeida

Doutora em Lingüística de Língua Portuguesa. Professora UFSCar, SP, Brasil