Os tipos de anotações, a codificação, e as interfaces do Projeto Lácio-Web: Quão longe estamos dos padrões internacionais para córpus?

Sandra Maria Aluísio[1,2], Leandro H. M. de Oliveira[1], Gisele Montilha Pinheiro[1]

[1] Núcleo Interinstitucional de Lingüística Computacional (NILC), CP 668, 13560-970 São Carlos, SP, Brasil

[2]ICMC-Universidade de São Paulo, CP 668, 13560-970 São Carlos, SP, Brasil

sandra@icmc.usp.br, {leandroh, gisele}@nilc.icmc.usp.br

Abstract *This paper addresses issues related to the development and standards for public available corpora, including types of annotation, encoding (i.e. forms of representation), tools and database architectures, in connection with the Lacio-Web Project. We also assess whether the decisions made for the Lacio-Web Project conform with the standards and how far are we from a representative Corpus of the Brazilian Portuguese.*

Resumo. *Neste artigo discutimos questões relacionadas aos tipos de anotação, codificação (no sentido de "forma de representação"), ferramentas e arquiteturas para dados, considerados padrões em ambientes de desenvolvimento e disponibilização de córpus. É discutido, também, quão próximas estão as decisões do Projeto Lácio-Web desses padrões e da construção de um Corpus Nacional do Português Brasileiro.*

1. Introdução

Vários córpus foram construídos para a língua inglesa, desde o pioneiro córpus Brown, lançado em 1964 com 1 milhão de ocorrências. Em termos de megacórpus balanceados, tanto o British National Corpus (BNC), para a variante britânica, quanto o American National Corpus (ANC), para a americana, contribuem para o desenvolvimento de ferramentas de processamento de língua natural (PLN) e para a descrição da língua e construção de recursos, tais como dicionários e gramáticas. Além disso, esses córpus impulsionam o desenvolvimento de formatos padrões de anotação e codificação, além de arquiteturas para dados e para ferramentas de manipulação de córpus. São esses padrões internacionais que ajudam a criar grandes córpus que sejam intensivamente usados, reusáveis e extensíveis.

Em [Ide and Brew 2000], a **reusabilidade** (característica de um córpus ser usável em mais de um projeto de pesquisa e por mais de um grupo de pesquisadores) e a **extensibilidade** (isto é, a capacidade de córpus serem melhorados em várias direções, por exemplo, com a provisão de um nível a mais de análise lingüística) são colocadas como dois aspectos a serem considerados em projetos de córpus. Para criar um Córpus Nacional do

Português Brasileiro (CNPB), objetivo de vários pesquisadores no Brasil[1], espera-se que tal megacórpus contemple uma **boa variedade de gêneros, tipos de textos** e **domínios** do conhecimento, inserida numa tipologia textual criteriosa e explícita. Também é desejável que o megacórpus seja **sincrônico** e **contemporâneo** como outros córpus desse tipo, trazendo a produção tanto escrita quanto falada em escala nacional. O córpus deve conter **textos completos** (escritos e transcritos), pois isso viabiliza um tipo especial de estudo, a análise do discurso e do texto. Há, entretanto, uma necessidade que não é de ordem técnica, mas que precede todas as outras, caso esse córpus envolva a disponibilização pública e integral via Web: a obtenção da **autorização de uso dos textos** para pesquisa.

Este artigo apresenta um projeto de desenvolvimento de corpus, o Lácio-Web (LW) [2] [Aluisio et al 2004, Aluisio et al 2003a, Aluisio et al 2003b], em direção à construção de um CNPB. Através do Projeto LW: a) propusemos uma tipologia ortogonal de textos, que privilegia criteriosamente o gênero e o tipo de texto, o domínio e o meio de distribuição; b) obtivemos a autorização de uso dos textos, possibilitando acesso livre desse material via Web; c) criamos uma interface Web de pesquisa e montagem de subcórpus, de modo a atender a maioria dos dados armazenados no cabeçalho das amostras; d) associamos a cada córpus (o LW possui seis tipos diferentes de córpus) um conjunto de ferramentas de processamento lingüístico, muitas das quais já utilizadas em outros projetos do Núcleo Interinstitucional de Lingüística Computacional (NILC)[3]; e e) adequamos o acesso aos córpus, a fim de torná-los de fácil interação entre os usuários especialista e leigos.

Na próxima seção apresentamos o Projeto Lácio-Web, seu status atual e o montante de dados e ferramentas a serem disponibilizados até o final do projeto. Na seção 3, são comentadas as vantagens do uso de XML para criação e manipulação de córpus e de padrões internacionais para codificação e intercâmbio de dados, com vistas à construção de um CNPB. Nessa seção também apresentamos as diferenças e semelhanças desses padrões com as decisões do LW. Na Seção 4, apresentamos as interfaces de pesquisa e de ferramentas.

2. O Projeto Lácio-Web (LW)

LW é um projeto iniciado em 2002, com 30 meses de duração, financiado pelo CNPq, e desenvolvido na Universidade de São Paulo pelo NILC, Instituto de Matemática e Estatística (IME)[4] e Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH)[5]. O Projeto visa ao desenvolvimento de vários tipos de córpus e ferramentas tanto para análise qualitativa (i.e., os dados podem ser utilizados, por exemplo, na construção de dicionários gerais ou terminológicos, ou ainda, na descrição da língua) quanto para a quantitativa (i.e., as estatísticas sobre os dados podem ser utilizadas, por exemplo, na construção de dicionários, etiquetadores morfossintáticos, sintáticos e corretores gramaticais). Os córpus do LW e suas ferramentas (Seção 4) são disponibilizados a partir de uma interface Web. Com respeito aos córpus, o Projeto LW traz: 1) um córpus aberto, sincrônico e contemporâneo de português escrito do Brasil (Lácio-Ref); 2) um córpus fechado, manualmente anotado com etiquetas morfossintáticas (Mac-Morpho); 3) um córpus fechado automaticamente anotado com lemas, etiquetas morfossintáticas e sintáticas para o qual será

[1] Veja em http://www.nilc.icmc.usp.br/iiiencontro/iiiencontro.htm as decisões do III Encontro de Córpus, realizado em 7 de novembro de 2003, no IEL, Unicamp.
[2] http://www.nilc.icmc.usp.br/lacioweb/
[3] http://www.nilc.icmc.usp.br/nilc/index.html
[4] http://www.ime.usp.br/
[5] http://www.fflch.usp.br

usado um *parser* desenvolvido no NILC[6] (Lácio-Sint); 4) um córpus aberto de desvio, contendo textos não revisados segundo os padrões da norma culta (Lácio-Dev); 5) um córpus paralelo aberto contendo textos em inglês e português do Brasil (Par-C), e 6) córpus comparáveis gerados automaticamente a partir de textos do Lácio-Ref e Ref-Ig (um córpus de referência do inglês construído no LW que traz, atualmente, textos originais em inglês do gênero jurídico) (Comp-C). Uma característica que distingue os córpus do LW com outros do português brasileiro é a sua proposta de servirem de *benchmark* para avaliar ferramentas de PLN. É o caso do Mac-Morpho para avaliação de etiquetadores morfossintáticos[7], Comp-C e partes do Lácio-Ref para avaliar métodos automáticos de extração de termos, Par-C para avaliação de alinhadores automáticos e Lácio-Dev para avaliação de corretores gramaticais.

O primeiro lançamento do Projeto se deu em 20/1/2004 e tornou dois córpus disponíveis que são detalhados abaixo: uma versão do Lácio-Ref para pesquisa e geração de subcórpus e o MAC-Morpho para download. O acesso aos córpus se dá após preenchimento de um cadastro.

A versão do Lácio-Ref possui 4.156.816 ocorrências, composta de textos organizados em cinco gêneros (Informativo, Científico, Prosa, Poesia e Drama), vários tipos de textos, vários domínios e alguns meios de distribuição (revista, internet, livro)[8]. O Lácio-Ref é disponibilizado para pesquisa com geração de subcórpus para download e acessado em dois formatos: a) texto com cabeçalho em XML, contendo dados bibliográficos e de classificação textual; e b) texto cru acrescido dos dados relativos ao título e à autoria. Nos subcórpus podem, ainda, ser aplicados três tipos de ferramentas: contadores de freqüência, concordanciadores e etiquetadores.

O MAC-Morpho possui 1.167.183 ocorrências de textos jornalísticos de dez cadernos da Folha de São Paulo, 1994. Essas foram etiquetadas pelo parser Palavras de Eckhard Bick (http://visl.hum.sdu.dk), mapeadas para o conjunto de etiquetas do Projeto Lácio-Web[9] e revisadas manualmente quanto à anotação morfossintática. O MAC-Morpho é disponibilizado para download em 2 formatos: um adequado para pesquisas lingüísticas com o uso de contadores de freqüência ou concordanceadores, por exemplo, e outro adequado ao treinamento de etiquetadores.

Para o lançamento final em junho, que culmina com o fim do suporte financeiro do CNPq, o Lácio-Ref será enriquecido com textos dos gêneros Instrucional, Jurídico, Informativo e Cientifico e contemplará muitos outros tipos de textos, domínios e meios de distribuição, totalizando 8.291.818 ocorrências. Como a tipologia prevê 9 gêneros, os dois restantes (Técnico-Administrativo e De Referência) serão contemplados em projetos de continuação do LW. Também haverá a disponibilização do córpus paralelo Par-C com 646 arquivos de textos em inglês e 646 em português da Revista Pesquisa Fapesp, totalizando 893.283 ocorrências e o lançamento da ferramenta de montagem de córpus comparáveis inglês-português envolvendo o gênero jurídico. Para a construção de córpus comparáveis, foi construído um córpus de referência de textos em inglês (Ref-Ig) do domínio jurídico. Ele

[6] http://www.nilc.icmc.usp.br/nilc/tools/curupira.html
[7] Três etiquetadores morfossintáticos disponíveis na Web foram treinados com o Mac-Morpho, podendo ser utilizados através de uma interface Web no LW. Veja as precisões de cada um em http://www.nilc.icmc.usp.br/lacioweb/ferramentas.htm
[8] Veja os textos do primeiro lançamento, separados por gênero, tipos de texto, domínios e meios de distribuição em http://www.nilc.icmc.usp.br/lacioweb/plancamento.htm
[9] Para saber mais sobre o processo de mapeamento, cf. http://www.nilc.icmc.usp.br/lacioweb/manuais

conta com 15 textos e 22.948 ocorrências e, futuramente, será ampliado. Os córpus Lácio-Dev e Lácio-Sint serão disponibilizados futuramente, como frutos de pesquisas de doutorado e mestrado, respectivamente. No total, o Projeto LW possuirá, no seu segundo lançamento, 5694 arquivos, totalizando 10.375.323 ocorrências.

3. Padrões internacionais para criação e manipulação de córpus

Discutiremos as questões sobre quais os **tipos de anotação**, **codificação** e **ferramentas e arquiteturas para ferramentas e dados**. Os dois primeiros estão bem descritos em [Ide and Romary 2003, Ide et al. 2003] e serão brevemente explicados aqui antes de explorarmos nossas opções. Por sua vez, as ferramentas disponíveis dependem da escolha da representação escolhida e deveriam ser livremente disponíveis e reusáveis para evitar o processo caro de reimplementação de software a cada novo projeto de córpus [Ide and Brew, 2000]; elas também serão exploradas nesse artigo.

Geralmente, distingue-se a anotação de segmentação da anotação lingüística. Na **anotação de segmentação** do texto cru, tem-se: 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, datas e ênfase. Já na **anotação lingüística** é fornecida a informação lingüística sobre segmentos como etiquetação morfossintática e sintática.

A **representação** se refere ao formato escolhido para explicitar a anotação. É aconselhável que a representação permita a separação entre os dados originais (ou também anotados com a estrutura geral) e anotações para que, por exemplo, possamos aplicar vários tipos de etiquetadores morfossintáticos ou sintáticos num mesmo córpus. Essa estratégia para a arquitetura dos dados que tem sido utilizada em projetos atuais de córpus [Ide and Macleod 2001, Santos and Bick 2000] é diferente da estratégia clássica de adicionar incrementalmente anotações aos dados originais. Mais detalhes dessa discussão estão em [Ide 1998, Ide and Romary 2003].

Um formato bastante usado e que provavelmente será o escolhido para representar a maioria dos córpus é a eXtensible Markup Language (XML) – um padrão internacional para representação e intercâmbio de dados na Web –, pois tem características e extensões úteis para a criação e manipulação de córpus anotados, entre elas: a) XML Links, que permitem endereçar os elementos XML tanto dentro de um mesmo documento como em outros documentos; b) a linguagem XPath e XPoint que, através de predicados, permitem localizar elementos na estrutura de elementos (em árvore) e selecionar fragmentos do texto; c) XSLT, que pode ser usada para converter um documento XML em outro formato; e d) *XML schemas* que estendem o poder dos DTD´s permitindo uma avaliação melhor tanto da forma quanto do conteúdo dos documentos XML [Ide 2000, Ide et al 2000]. XML não é, porém, o único formato para codificar córpus. O IMS Corpus Workbench[10] foi usado no Projeto AC/DC [Santos and Sarmento 2003, Santos and Bick 2000] para disponibilizar córpus do português europeu e brasileiro, e no Projeto Korpus 2000 [Andersen et al 2002], para o dinamarquês.

É interessante contrastar a abordagem gerencial de criação de grandes córpus realizada no projeto AC/DC com uma outra para a criação de um CNPB se decidirmos

[10] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

utilizar um padrão internacional para codificação de córpus como o Corpus Encoding Standard (CES)[11]. O CES é uma aplicação do SGML e possui uma versão mais atual em XML, o XCES. O CES utiliza e adapta os padrões das diretrizes para codificação e intercâmbio de textos eletrônicos TEI[12] para codificar córpus. Na medida em que a abordagem utilizada no projeto AC/DC harmoniza a anotação de segmentação de vários córpus e os centraliza num único site para que os mesmos possam ser pesquisados, a abordagem XCES pretende um desenvolvimento distribuído de anotações e ferramentas de acesso a córpus (uma vez que esse padrão seja adotado por vários projetos de córpus dispersos geograficamente), podendo os dados, anotações e ferramentas de pesquisa serem armazenados em vários servidores dispersos geograficamente. Essa última abordagem favorece a construção de um CNPB, dado que o volume enorme de dados envolvidos e o trabalho de construção de tal recurso inviabilizam que o mesmo seja construído por um único grupo de pesquisa. Além disso, permite utilizar os vários córpus escritos e de fala já construídos. É importante notar, entretanto, que a tarefa de construção de um CNPB não é trivial, envolve altos recursos financeiros e recursos humanos treinados, além de tempo. Essa empreitada, entretanto, faz com que a comunidade de lingüística de córpus envolvida na construção de um CNPB esteja afinada com os padrões de disponibilização de recursos internacionais. O córpus Lácio-Ref e sua tipologia de textos podem fazer parte de um futuro CNPB, entretanto há que se cuidar do balanceamento de gêneros. Foi também uma decisão de projeto privilegiar textos integrais e, assim, o Lácio-Ref não se conforma com as decisões de córpus como o BNC que limita o tamanho das amostras de textos. Tal mudança, contudo, pode ser facilmente realizada.

Várias decisões no projeto do LW ainda estão distantes dos padrões internacionais (como o XCES) tanto com relação à anotação como à codificação. No LW, não há a anotação de grande parte dos elementos da estrutura geral, tais como capítulos, parágrafos, subparágrafos, títulos e notas de rodapé em XML. Porém, eles estão formatados e padronizados para fácil visualização com marcas do tipo quebra de linhas, caixa alta, etc.. Já os elementos gráficos estão todos anotados em XML. No córpus Mac-Morpho, a anotação morfossintática se dá num formato propício para o treinamento de etiquetadores (cada palavra em uma linha juntamente com sua etiqueta); nenhum esforço para separação entre texto cru e anotação foi realizado. Quanto à anotação dos elementos do cabeçalho, esses estão em XML e podem facilmente se adequar às normas do XCES. Um editor de cabeçalho certifica que a geração desse esteja correta e facilita a edição das várias informações. Quanto ao grande trabalho de reescrita de ferramentas de PLN e manipulação de córpus em geral, ele pode ser evitado com a tendência atual de utilizar arquiteturas para construção e manipulação de córpus como GATE[13]. Na abordagem do LW foram colocadas à disposição ferramentas desenvolvidas em projetos realizados durante os 10 anos do NILC, como etiquetadores morfossintáticos, sintáticos, alinhadores automáticos de sentenças e extratores de termos. Assim, privilegia-se o reuso de software.

4. Interfaces de pesquisa e de ferramentas do Portal LW

Uma vez que os principais objetivos do Lácio-Web são a disponibilização de córpus e de ferramentas para análise lingüística e ferramentas de PLN, foram desenvolvidos dois tipos

[11] http://www.cs.vassar.edu/CES/
[12] http://www.tei-c.org/
[13] http://gate.ac.uk

de interfaces: as interfaces de pesquisa de córpus e as interfaces de ferramentas. Apesar de serem interligadas e interdependentes, os objetivos são diferentes. As **interfaces de pesquisa** têm as funções de: i) possibilitar a pesquisa de córpus obedecendo aos critérios de classificação tipológica e bibliográfica dos textos; e ii) com o resultado das pesquisas, promover a montagem de subcórpus. Já as **interfaces de ferramentas** têm como objetivo aplicar as ferramentas disponíveis no LW nos córpus ou subcórpus montados pelos usuários e, conseqüentemente, exibir os resultados. A principal motivação para a criação destas interfaces foi o fato de se garantir o direito dos usuários (especialistas ou não) de pesquisar córpus, bem como o de montar seus subcórpus e sobre eles aplicar, de forma independente, as ferramentas disponíveis.

As interfaces de pesquisa e de ferramentas que discutiremos neste artigo são referentes ao córpus Lácio-Ref e MAC-Morpho, no escopo do LW. Todos os textos pertencentes aos córpus do LW são classificados quanto a dois conjuntos de informações: 1) informações bibliográficas (dados de catalogação: amostragem, título, fonte, status de língua, autoria, tradução, etc.) e 2) informações tipológicas (dados de classificação: gênero e subgênero textual, tipo de texto, domínio e subdomínio, meio de distribuição). Essas informações são únicas e exclusivas de cada texto e são armazenadas num cabeçalho descrito em linguagem XML[14].

A pesquisa dos córpus disponíveis no Portal do LW é depende dessas informações, visto que os campos disponíveis nas interfaces de pesquisa advêm do cabeçalho. Os dados de classificação obedecem a uma hierarquia interna que dita uma relação de subordinação entre os campos (própria da codificação XML), permitindo uma coerência de classificação do texto. Assim como esses, os dados de catalogação também possuem tal relação. Um exemplo: se determinado texto pertence ao domínio Ciências Exatas e da Terra, significa que o subdomínio do texto deverá ser subdomínio subjacente a este domínio, e assim sucessivamente em todas classificações disponíveis no cabeçalho.

Para garantir a execução rápida das consultas dos usuários e a atualização dinâmica da interface, considerando as relações hierárquicas dos campos do cabeçalho, foi necessária a transposição da estrutura do cabeçalho XML numa estrutura de banco de dados relacionais. Essa atualização dinâmica da interface diz respeito à sensibilidade e flexibilidade dos campos de seleção (para pesquisa) pertencentes às interfaces e, conseqüentemente, coerentes com a classificação dos textos. Isso quer dizer que os campos de seleção mostrados na interface são sensíveis ao conteúdo selecionado pelo usuário num dado momento, e podem alterar dinamicamente o conteúdo dos outros campos da pesquisa. A adoção do banco de dados também foi importante para garantir a eficiência e a rapidez no processamento das pesquisas (consultas) no ambiente Web, visto que tais pesquisas exigem a aplicação de vários *joints* (junção de dados relacionados) de tabelas, implementados por meio do uso da linguagem *SQL (Structured Query Language)*. Mencione-se, também, que a utilização de um banco de dados relacional aumenta a segurança e a integridade dos dados armazenados.

---

[14] É importante assinalar alguns pontos aqui pertinentes: a) o Mac-Morpho não tem, até momento, o c-LW inserido nas amostras; b) o tratamento de outros córpus pode gerar a criação de novos dados de catalogação; e c) cada uma das categorias da catalogação possui desdobramentos que se constituem características importantes sobre textos escritos (tipo de autoria, data e local de publicação, link dos dados de tradução para o texto original, etc.).

4.1. As interfaces de pesquisa

Foram definidos três tipos de pesquisas, cujos critérios se nortearam pela expectativa dos tipos de usuários de córpus. De um lado estão os usuários **especialistas,** como lingüistas, gramáticos, analistas do texto e do discurso, sociolingüistas, teóricos da literatura, lingüistas computacionais, lingüistas de córpus, lexicógrafos, terminólogos, cientistas da computação. De outro, os usuários **leigos**: estudantes de toda sorte, revisores de texto, professores de língua, tradutores, historiadores, etc.. À disposição desse público-alvo foram projetadas as seguintes opções de seleção de subcórpus: pesquisa simples, a pesquisa avançada e a pesquisa personalizada.

4.1.1 Pesquisa Simples: é a mais genérica do Portal e, ao mesmo tempo, a que oferece menos opções de seleção aos usuários. O seu caráter genérico se define pela vinculação do sistema de busca com o padrão de nomeação dos arquivos, em que se prevê a seleção de subcórpus pela escolha dos dados relativos à classificação das amostras textuais. Por sua vez, o parâmetro da nomeação de arquivos busca atender às expectativas de um usuário leigo para quem a pesquisa avançada e/ou personalizada podem indicar dificuldade ou não-relevância. Os resultados obtidos (i.e., número de textos recuperados) pela pesquisa simples são, geralmente, extensos.

4.1.2 Pesquisa Avançada: é a intermediária, situada entre a Simples e a Personalizada, permitindo que o usuário refine suas opções e obtenha resultados mais específicos, mas em menor grau que a busca personalizada. Deixa de ser relacionada à nomeação dos arquivos e passa a disponibilizar os dados da catalogação na seleção de subcórpus pelo usuário. Nesse caso, o sujeito que se espera no acesso é o usuário que precisa refinar o seu subcórpus em termos mais definidos de amostras e que é capaz de julgar os textos em termos mais específicos de classificação. Por exemplo, é capaz de dizer que quer apenas textos literários em prosa – biografia, respectivamente o gênero e o subgênero textual. Assim, além dos campos da Pesquisa Simples, os usuários podem selecionar mais dados de classificação (*Supergênero, Gênero* e *Subgênero textual*), bem como os dados de catalogação bibliográficas, como *Nome de Autor*, *Nome do Periódico* e *Caderno*. Os campos *Gênero*, *Subgênero*, *Nome do Periódico* e *Caderno* também possuem conteúdos dinâmicos. A Figura 1 mostra duas telas (A e B) como exemplo deste tipo de seleção.

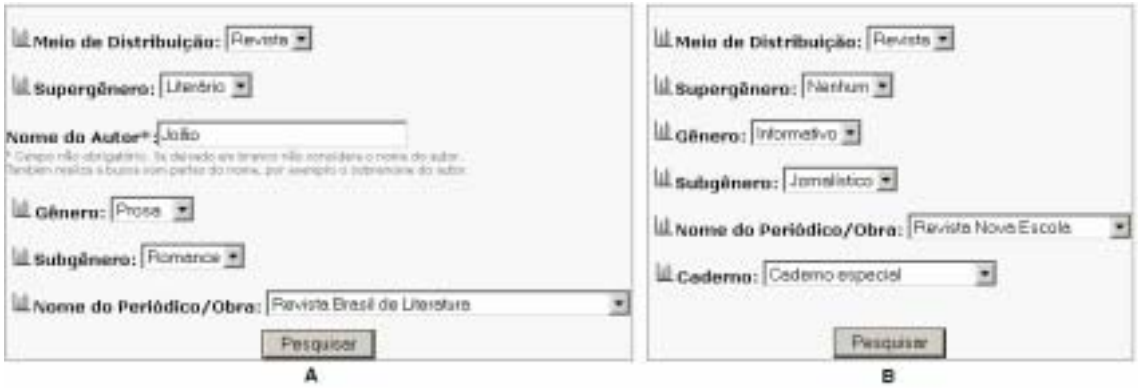

Figura 1 – Telas da Pesquisa Avançada no Portal LW

Observe que nessa ilustração o campo *Caderno* não aparece como opção disponível. Isso acontece porque o *Nome do Periódico* selecionado – a "Revista Brasil de Literatura" – não possui cadernos vinculados. Entretanto, quando o *Supergênero* selecionado é "Literário", o campo *Nome de Autor* é ativado. Em contrapartida, observando a Figura 1-B,

que representa outro exemplo da *Pesquisa Avançada*, verificamos que não aparece o campo *Nome do Autor*; desta vez, o campo *Caderno* está disponível visto que o *Nome do Periódico* "Revista Nova Escola" possui cadernos vinculados.

4.1.3 Pesquisa Personalizada: permite ao usuário refinar sua pesquisa ao máximo, oferecendo opções de seleção que abrigam, em dois grupos, tanto os dados de catalogação como os de classificação. Foi projetada para o usuário especialista, que recorta criteriosamente suas amostras e está a par de todos os detalhes de publicação dos textos que procura. Nessa pesquisa o usuário deve definir detalhadamente o recorte de sua investigação, de maneira que os resultados obtidos sejam de um perfil específico. Novos campos de seleção como: o *Tipo de Amostragem,* o *Tamanho da Amostra*, o *Tipo de Autoria* e o *Tipo Textual* são apresentados ao usuário, sendo que a grande maioria deles possui conteúdos dinâmicos. Como exemplos dessa dinamicidade estão os "novos" campos de *Domínio* e *Subdomínio,* cujos conteúdos são dependentes.

4.2. As interfaces de ferramentas

As interfaces de ferramentas do Portal do LW têm como principal objetivo facilitar a aplicação de ferramentas de análise lingüística aos córpus e/ou subcórpus montados pelos usuários. Sua maior vantagem é a condução do usuário na tarefa de verificar, por meio de ferramentas, a qualidade e relevância dos subcórpus montados. Atualmente, há quatro ferramentas disponíveis no LW. Três delas são aplicadas ao córpus Lácio-Ref e, conseqüentemente, aos subcórpus montados pelos usuários ((a), (b) e (c) abaixo). A outra (um concordanceador) é especificamente aplicada ao córpus etiquetado MAC-Morpho.

a) Contador de Freqüência Padrão: calcula a freqüência com que as palavras ocorrem em um córpus, ferramenta comum em trabalhos com córpus, já que calcular a freqüência de palavras é uma tarefa simples. Porém, o contador de freqüência disponível no Portal LW possui um diferencial relevante, que é o reconhecimento de "lexias complexas dos nomes próprios"[15] e "palavras compostas"[16] para o cálculo das freqüências. Nesse contador, o reconhecimento dos tokens é realizado por um conjunto de regras de formação de palavras ao qual o contador é submetido no momento de sua execução. Além disso, no contador os usuários têm a opção de escolher a ordem de freqüência das palavras (alfabética ou decrescente) e também em qual córpus deseja aplicar o contador. O resultado do contador de freqüência padrão traz diversas informações a respeito das palavras ou expressões do córpus: i) a quantidade de textos pertencentes ao córpus; ii) a quantidade de ocorrências simples (tokens) do córpus; iii) a quantidade de ocorrências simples que aparecem apenas uma vez, bem como, as que aparecem mais de uma vez; iv) a quantidade de "palavras" (lexias complexas) que aparecem no córpus; v) a quantidade de "palavras" que aparecem apenas uma vez, bem como as que aparecem mais de uma vez, e finalmente; e vi) o índice vocabular, que indica a variedade do vocabulário utilizado no córpus.

b) Contador de Freqüência por Palavra ou Expressão: possui a funcionalidade de contar a freqüência de uma palavra ou expressão previamente fornecida pelo usuário. Esse contador é semelhante ao descrito anteriormente, mas, nesse caso, uma palavra ou expressão é requerida como entrada. A palavra ou expressão fornecida pelo usuário pode ser também

---

[15] *Lexia complexa* pode ser entendido como a unidade de significação composta de mais de um token não unidas por meio de hífen. Ex.: ticket refeição, vale transporte, virgem Maria, etc..

[16] Aqui, considera-se palavra complexa as unidades de significação unidas por meio de hífen ou que se constituem pela união de uma seqüência alfabética e outra numérica. Ex.: sem-terra, pára-choque, Largo 13, Pio XI, etc.

uma palavra composta ou lexia complexa. São também dados de entrada o córpus de origem e a "janela" (quantidade de palavras no contexto superior e inferior) onde a mesma aparece.

c) Concordaceador: essa ferramenta tem o objetivo de destacar uma determinada palavra ou expressão no texto onde ela ocorre. O concordanceador implementado no Portal LW possui várias opções que podem ser definidas pelos usuários. Por exemplo, a definição do tamanho do contexto (reduzido e expandido) que dizem respeito, respectivamente, ao tamanho (em caracteres) do segmento e do parágrafo onde a palavra ou expressão aparece, bem como o "nível de sensibilidade", que pode ser: *Igual a, Começando com, Terminando com* e *Contendo* dos mesmos a serem pesquisados no córpus. O resultado da aplicação dessa ferramenta traz todos os trechos (contexto reduzido) nos quais a palavra aparece, sendo que um link sobre a palavra "alvo" leva o usuário aos contextos expandidos. Um concordanceador semelhante é aplicado no córpus Mac-Morpho, com a diferença de que o usuário pode também definir qual etiqueta da palavra ou expressão ele deseja considerar.

Uma importante vantagem das interfaces de ferramentas descritas nesta seção é que todos os resultados podem ser salvos pelo usuário através do link "download do resultado", disponibilizado pela interface. Esta característica oferece maior flexibilidade de navegação e uso de córpus aos usuários, visto que, uma vez aplicadas as ferramentas, os usuários podem, além de se envolver criteriosamente na escolha de seu subcórpus, salvar os seus resultados localmente, o que permite analisá-los posteriormente, i.e., fora do ambiente do portal.

5. Conclusões e Trabalhos Futuros

Quase ao final de 30 meses de pesquisa e desenvolvimento, o LW disponibiliza, de forma gratuita: a) 4 tipos distintos de córpus (Lácio-Ref, Mac-Morpho, Par-C e Comp-C); b) ferramentas de processamento lingüístico-computacional (contador de freqüência, concordanceador e etiquetadores morfossintáticos); e c) Portal com 3 tipos de interface de pesquisa, com ferramentas de base associadas. É, também, um ambiente de navegação dinâmica, didática e, sobretudo, de incentivo ao uso de córpus para os mais diversos tipos de investigação lingüística, uma vez que permite o download completo das amostras dos córpus. É um primeiro passo para um trabalho conjunto de construção de um CNPB.

Embora várias decisões tomadas no projeto do LW ainda estão um pouco distantes dos padrões internacionais (como o XCES) tanto com relação à anotação como à codificação, demos 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. Num possível retorno ao Projeto, espera-se que as limitações na construção e disponibilização de córpus sejam eliminadas: preenchimento com amostras textuais das categorias de gênero e tipo textual, domínio e meio de distribuição não contempladas; estudo e aplicação do balanceamento de córpus; refinamento de ferramentas; associação de novas ferramentas aos córpus e/ou ferramentas já existentes a outros córpus.

Referências

Aluísio, S. M., Pinheiro, G. M., Finger, M., Nunes, M.G.V. and Tagnin, S. E. O. (2003a) "The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation", Corpus Linguistics 2003, Lancaster, UK, Proceedings of Corpus Linguistics 2003. Lancaster: 2003. v. 16, 14-21.

Aluísio, S. M., Pelizzoni, J. M., Marchi, A. R., Oliveira, L. H., Manenti, R. and Maquiafável, V. (2003b) “An account of the challenge of tagging a reference corpus of Brazilian Portuguese”, Lecture Notes on Artificial Intelligence 2721, 110-117.

Aluísio, S. M., Pinheiro, G. M., Manfrim, A. M. P., Oliveira, L. H. M. de, Genovês 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”, LREC 2004. Proceedings of LREC, 2004, Lisboa, Portugal.

Andersen, M. S., Asmussen, H. e Asmussen, J. (2002) “The project of Korpus 2000 going public”, Proceedings of Euralex 2002, 291-299.

Ide, N. e Romary, L. (2003). “Outline of the International Standard Linguistic Annotation Framework.”, Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right, Sapporo, 1-5.

Ide, N., Romary, L., de la Clergerie, E. (2003). “International Standard for a Linguistic Annotation Framework”, Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology, Edmunton.

Ide, N. e Macleod, C. (2001). “The American National Corpus: A Standardized Resource of American English”, Proceedings of Corpus Linguistics 2001, Lancaster UK.

Ide, N. (2000). “The XML Framework and Its Implications for the Development of Natural Language Processing Tools”, Proceedings of the COLING Workshop on Using Toolsets and Architectures to Build NLP Systems, Luxembourg, 5 August 2000.

Ide, N. e Brew, C. (2000). “Requirements, Tools, and Architectures for Annotated Corpora”, Proceedings of Data Architectures and Software Support for Large Corpora. Paris: European Language Resources Association, 1-5.

Ide, N., Bonhomme, P. e Romary, L. (2000). “XCES: An XML-based Standard for Linguistic Corpora”, Proceedings of the Second Language Resources and Evaluation Conference (LREC), Athens, Greece, 825-830.

Ide, N. (1998). “Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora”, Proceedings of the First International Language Resources and Evaluation Conference, Granada, Spain, 463-470.

Santos D. e Sarmento, L. (2003). "O projecto AC/DC: acesso a corpora / disponibilização de corpora", Amália Mendes & Tiago Freitas (orgs.), Anais do XVIII Encontro da Associação Portuguesa de Linguística (Porto, 2-4 de Outubro de 2002), APL, 2003, 705-717.

Santos, D. e Bick, E. (2000) “Providing Internet Acces to Portuguese Corpora: the AC/DC Project”, Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), 205-210. Atenas, 31 May-2 June 2000.