CoMAprend – a experiência da construção de um *corpus* de aprendizes para estudos

Stella Esther Ortweiler Tagnin[*]
Guilherme Fromm[**]

Resumo: O objetivo deste artigo é mostrar o processo de construção de um *site* na Internet especialmente criado para reunir textos escritos por alunos dos cursos de línguas inglesa, francesa, alemã, italiana e espanhola numa universidade brasileira para formar um *corpus* (ou *corpora*) para pesquisadores. Os aprendizes inserem suas redações no site e o professor/pesquisador pode, então, construir um *subcorpus* de aprendizes (de acordo com a língua, por exemplo) para estudos específicos e pesquisar o mesmo com ferramentas de análise disponíveis no *site* (gerador de listas de palavras, concordanciador, gerador de n-gramas) ou exportá-lo para investigação com outras ferramentas de análise lexical *off-line*.
Palavras-chave: Aprendizes de Línguas, Linguística de Corpus, Corpora de Aprendizes

Abstract: This article aims at showing the process of constructing an Internet site specially designed to collect texts written by language learners of English, French, German, Italian and Spanish at a Brazilian university to form a corpus (or corpora) for research. Learners populated the site with their written assignments and the teacher/researcher can then build a learner subcorpus (according to language, for example) for specific studies and query this subcorpus with built-in tools (wordlist generator, concordancer, n-grams generator) or export it for investigation with other stand-alone lexical analysis tools.
Keywords: Language Learners, Corpus Linguistics, Learner Corpora.

Quantos professores não guardam – ou teriam vontade de guardar – as redações de seus alunos para desenvolver futuras pesquisas? Mas, em geral, essa intenção não é levada a cabo, por vários motivos. O primeiro é a dificuldade de armazenamento: onde e como guardá-las? Mesmo que essa parte seja resolvida, restaria o problema de como usar o material para pesquisa, ou seja, como extrair desses textos escritos os itens a serem analisados, como comparar os textos, como identificar os erros mais comuns, para citar apenas alguns.

Na era digital, não faz mais sentido fazer pesquisas lingüísticas que não sejam com o auxílio de ferramentas eletrônicas. Para isso, no entanto, é preciso que os textos estejam digitalizados para poderem ser "lidos" por essas ferramentas.

A Linguística de Corpus, há já algum tempo, dedica-se a esse tipo de pesquisa e numerosas coletâneas de textos em formato eletrônico foram compiladas para os mais variados objetivos: corpora de língua geral como o Cobuild[1], o BNC[2], o COCA[3] para a

[*] Professora Associada do Departamento de Letras Modernas (Língua Inglesa) da FFLCH/USP.
[**] Doutor em Letras, Língua Inglesa, pela FFLCH/USP.
[1] Parte desse corpus pode ser acessado *online* em http://www.collins.co.uk/Corpus/CorpusSearch.aspx.
[2] British National Corpus, que pode ser consultado *online* em http://corpus.byu.edu/bnc/.
[3] Corpus of Contemporary American English, compilado por Mark Davies e disponbilizado em http://www.americancorpus.org/.

língua inglesa, e o Lácio-Web[4] para a língua portuguesa.; corpora de linguagem especializada, como os que compõem o CorTec do projeto COMET[5]; corpora paralelos (textos originais e suas respectivas traduções), como o COMPARA[6] e, mais recentemente, o CorTrad, também este parte do projeto COMET e, finalmente, os corpora de aprendizes.

Essa última variante surgiu no início da década de 1990 (Granger 1993a, 1993b, 1994), sob a direção de Sylviane Granger, da Universidade de Louvain, na Bélgica, e denominou-se ICLE (International Corpus of Learner English)[7]. O projeto continua em desenvolvimento e tem por objetivo coletar 200.000 palavras em redações do tipo argumentativo de estudantes de diversas nacionalidades aprendendo inglês. Na esteira desse projeto foram surgindo outros similares em vários países, ensejando numerosas pesquisas e publicações, entre as quais podemos citar Gilquin (2005) e coletâneas de artigos como Gilquin et alli (2008) e Granger & Meunier (2005).

Até há pouco, o único corpus de aprendizes em construção no Brasil era o Br-Icle[8] (Berber Sardinha 2001), a parte referente ao português brasileiro do projeto ICLE. Até o momento, o Br-Icle contém 80.000 palavras de textos argumentativos coletados pela Pontifícia Universidade Católica e algumas outras universidades que se associaram ao projeto. Numa segunda fase, recentemente lançada (2009), no entanto, o projeto ICLE não contou ainda com a língua portuguesa, variante brasileira.

Na Universidade de São Paulo, no entanto, foi criado um corpus de aprendizes – o CoMAprend (Tagnin, 2006) - com uma peculiaridade: abarca as redações dos aprendizes de língua estrangeiras das cinco áreas do Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas – alemão, espanhol, francês, inglês e italiano – , incluindo não só material produzido nos cursos de graduação, como também em seus cursos extracurriculares. Salvo engano, isso significa que, ao contrário dos corpora citados, em que temos aprendizes de diversas nacionalidades aprendendo a mesma língua, o inglês, no CoMAprend, temos uma nacionalidade comum aos aprendizes, a brasileira, e cinco línguas-alvo, ou seja, um corpus de aprendizes multilíngue.

[4] www.nilc.icmc.usp.br/lacioweb.
[5] www.fflch.usp.br/dlm/comet.
[6] Corpus bidirecional de textos literários em inglês e português www.linguateca.pt/COMPARA
[7] http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm
[8] http://www2.lael.pucsp.br/corpora/bricle/index.htm

Toda essa produção está concentrada num só banco de dados, o que permite uma variedade de pesquisas a partir de um único corpus. Com esse desenho o CoMAprend permite tanto pesquisas horizontais, como a comparação entre alunos da mesma classe ou de classes do mesmo nível, quanto pesquisas verticais ou diacrônicas, acompanhando o desenvolvimento de um aluno ou de todo um grupo (cf. Leech 1998, Kaszubski 2000, Lenko-Szymanska. 2000, entre outros).

Também é possível investigar estratégias de redação como a paráfrase, o sub- ou superuso de certas estruturas sintáticas, itens lexicais, colocações ou fórmulas (cf. Altenberg & Tapper 1998; Altenberg 2002, Berber Sardinha 2001, De Cock 1998, Granger, 1998a, 1998b dentre muitos outros)

Considerando que o subcorpus de cada língua já é um corpus comparável por abrigar redações provenientes de cursos com objetivos distintos – os de graduação e os extracurriculares –, outra área contrastiva que pode ser explorada é a avaliação da eficácia de diferentes métodos ou materiais usados na graduação e nos cursos extracurriculares.

De maior interesse talvez seja a possibilidade de desenvolver pesquisas translinguais, como verificar se determinadas dificuldades dos alunos são dependentes da língua estrangeira (LE) ou da língua materna (LM). Por exemplo, será que todos os aprendizes brasileiros têm problemas no aprendizado de preposições em qualquer das cinco línguas?Com o decorrer do tempo, outras universidades agregaram-se ao projeto, enriquecendo seu acervo, em especial para a língua inglesa.

Mas, nem tudo tem sido fácil nesse percurso. Talvez a maior dificuldade consista em motivar os professores – muitos deles ainda resistentes a novas tecnologias – a incentivarem os alunos para que enviem suas redações, ou seja, para que acessem o site do CoMAprend e lá copiem, no local indicado, sua redação. Para evitar problemas autorais, cada aluno deve preencher, uma vez por semestre, uma ficha com seus dados pessoais e uma autorização para que o material que venha a submeter possa ser usado para fins de pesquisa.

No intuito de compartilhar nossa experiência – quer para servir de parâmetro para outros pesquisadores, quer para evitar a repetição de nossos erros – passamos a relatar as etapas que levaram ao formato atual do CoMAprend.

O projeto inicial

Um *corpus*, segundo Tagnin (2005), é "[...] uma coletânea de textos, necessariamente em formato eletrônico, compilados e organizados segundo critérios ditados pelo objetivo da pesquisa a que se destina". A organização do *Corpus* Multilíngüe de Aprendizes (CoMAprend, disponível em http://www.fflch.usp.br/dlm/comet/comaprend.html), do Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, começou a ser discutida entre os integrantes do Projeto de *Corpus* Multilíngüe de Ensino e Tradução (COMET) e os representantes dos cursos de Letras Modernas (Alemão, Espanhol, Francês, Inglês e Italiano) em 2004. A idéia era reunir, num só lugar e em meio informatizado, a produção dos alunos de graduação (aulas de produção de texto) e de cursos livres (redações solicitadas pelos professores) ministrados no DLM da FFLCH. A concepção desse *corpus* ficaria, segundo a proposta de Berber Sardinha (2004, p. 20-21), assim configurado:

a. modo: escrito;
b. tempos: sincrônico/contemporâneo;
c. seleções: monitor (dinâmico, reciclável);
d. conteúdo: multilíngüe como um todo, monolíngüe em cada língua;
e. autoria: de aprendiz (falantes não-nativos);
f. finalidades: de estudo (*corpus* a ser descrito).

Ficou claro, desde o início, que o *corpus* a ser construído deveria ser, se não totalmente, pelo menos parcialmente, informatizado. Vários encontros foram promovidos, em que os representantes discutiram as especificidades de cada curso, tanto para os alunos de graduação quanto para os alunos dos cursos extracurriculares. Diversas metodologias de construção foram propostas e, a partir daí, elaborou-se um projeto para posterior confecção do banco de dados e do site de acesso ao qual ele estaria ligado. O primeiro documento, contemplando as necessidades para a elaboração do banco de dados e do site, previa:

1. a criação de um banco de dados para catalogar e gerar um código individual de alunos;

2. a construção de uma página na Internet onde os alunos pudessem, através de seu código individual, enviar suas redações;
3. o envio dessas redações, concomitantemente, para outro banco de dados que formaria o *corpus* e para o professor da disciplina. Deveria ser construído um banco de dados para cada uma das cinco línguas oferecidas pelo DLM e cada divisão (graduação ou curso no campus) dentro dessas línguas (variável, de acordo com a língua);
4. o desenvolvimento de um programa que selecionasse as redações do *corpus*, para fins de análise linguística e conforme as necessidades do pesquisador e possibilitasse a inserção de um cabeçalho;
5. a geração, em base semestral, de um CD-ROM com todos os *corpora* para ser disponibilizado aos participantes.

Foi sugerida, quando das discussões iniciais acerca do projeto, uma codificação para que o sistema gerasse um arquivo para cada redação inserida: o código IA40201, por exemplo, significaria que aquela redação pertenceria ao curso I (Inglês no Campus), ao ano A (2º semestre de 2003), o nível do curso ou a disciplina de graduação e o número da redação[9]:

IA400021

I	Língua/curso
A	Semestre/ano
4	Nível do curso[10]
0002	Código aluno
1	Número da redação

Essa proposta foi abandonada, por sugestão dos programadores, no momento da elaboração do banco de dados. Várias outras idéias também foram abandonadas e/ou alteradas durante a construção efetiva do banco de dados e do *site*. Muitas se mostraram ambiciosas demais, outras simples demais. A visão dos programadores foi essencial para as mudanças efetuadas.

Já com alguns pré-requisitos discutidos, o COMET, contando com o apoio de um financiamento do CNPq (400988/2006-2), orçou, em 2006, o valor para a

[9] Consideramos improvável a hipótese de que um aluno escrevesse mais de nove redações por semestre.
[10] Como, por exemplo, básico, intermediário e avançado.

construção do banco de dados e do site. Optamos pela Empresa Jr. do ICMC da USP de São Carlos. Depois de várias discussões em reuniões e troca de correspondência, foram decididos os requisitos finais do sistema[11] e a construção do mesmo. A página entrou no ar em setembro de 2005 e acaba de passar por uma reformulação (início de 2009) para automatizar e melhorar os processos de coleta de textos, estando agora em fase de testes.

A primeira fase do sistema

Entre a fase de elaboração do projeto inicial e o início das operações do CoMAprend, muitas alterações foram sugeridas pelos desenvolvedores, aprovadas pelo COMET e implementadas no banco de dados e na página de inserção de dados na Internet.

A inscrição dos alunos

Havia, antes da criação do projeto, questionários, em papel, que alguns cursos disponibilizavam para os alunos que concordassem em participar de projetos de elaboração de corpora. A idéia era informatizar esses questionários. Através de uma análise dos questionários já existentes, elaborou-se uma nova proposta e a mesma foi disponibilizada pelo site. Quando da inscrição do aluno, ele deve fornecer seus dados pessoais para o banco (a figura 1 mostra um mostra um formulário já preenchido); os mesmos dados são usados para a seleção de redações através de filtros. As informações a serem preenchidas são:

1. Nome;
2. sexo;
3. data de nascimento;
4. nacionalidade;
5. língua paterna/materna;
6. escolaridade: médio, superior, pós-graduação;
7. conhecimento de línguas estrangeiras;
8. anos de estudo de línguas estrangeiras;

[11] Elaborados pela Empresa Jr. de acordo com as discussões prévias entre todos os participantes.

9. residência no exterior;
10. língua(s) estrangeira(s) falada(s) em casa;
11. hábitos de contato com a língua estrangeira (selecionar em caixa de opção: internet, chat, filmes, livros, revistas, jornais, outros);
12. autorização para que as redações sejam usadas para fins acadêmicos.

Figura 1 - Dados de inscrição do aluno.

Além desses dados iniciais, a cada submissão de redação, o aluno deve completar outros campos (figura 2):

1. Tipo de redação, em menu drop-down: narração, descrição, argumentação, diálogo, informativo;
2. título da redação;
3. uso de material de referência, em menu drop-down: dicionário bilíngüe, dicionário monolíngüe, gramática, internet, bibliografia especializada, *corpora*, outros.

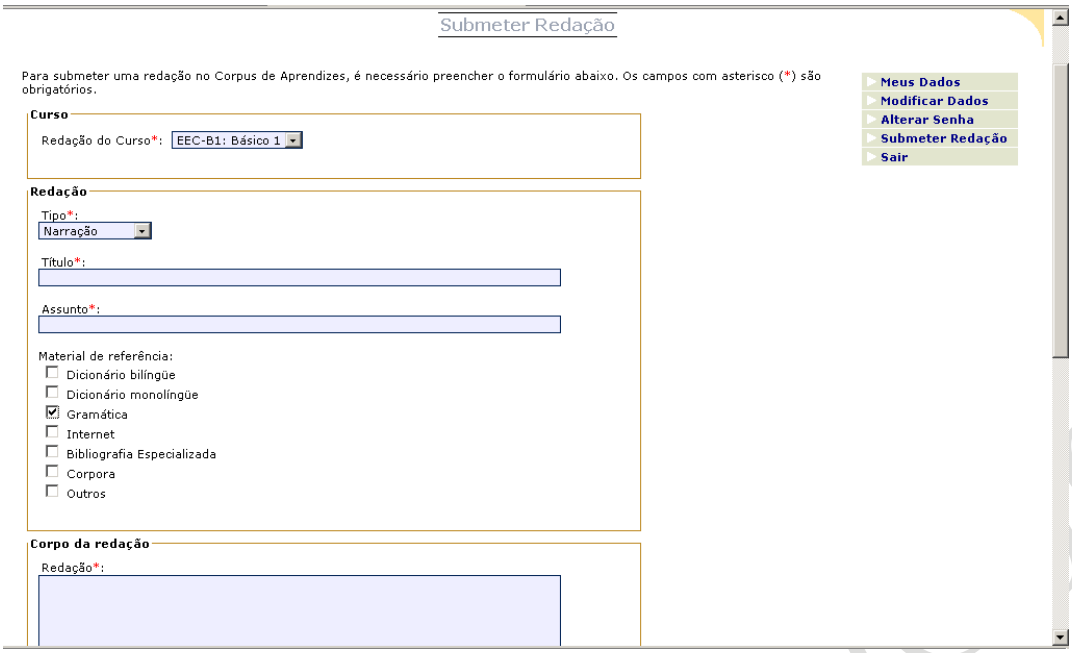

Figura 2 - Página de inserção da redação.

O sistema, por sua vez, gera automaticamente a data de submissão de cada redação.

O acesso dos pesquisadores

Qualquer pesquisador, ligado ou não à USP, ao solicitar uma senha para os coordenadores do projeto[12], recebe a permissão de acesso para pesquisar o site. A pesquisa é realizada através de vários filtros, relacionando os dados pessoais dos alunos (já apresentados acima) com as informações de turmas e cursos disponíveis.

Na página de pesquisa (fig. 3), vários modos são listados para buscas no banco de dados: informações pessoais dos alunos, por exemplo, pesquisar somente redações de alunos do sexo masculino, características de redações (pesquisa por data de submissão, tipo, assunto, título, referências, corpo do texto) e busca rápida pelos cursos disponíveis no semestre atual ou anteriores.

[12] Atualmente Stella E. O. Tagnin, Cristina A. E. Kindermann e Guilherme Fromm.

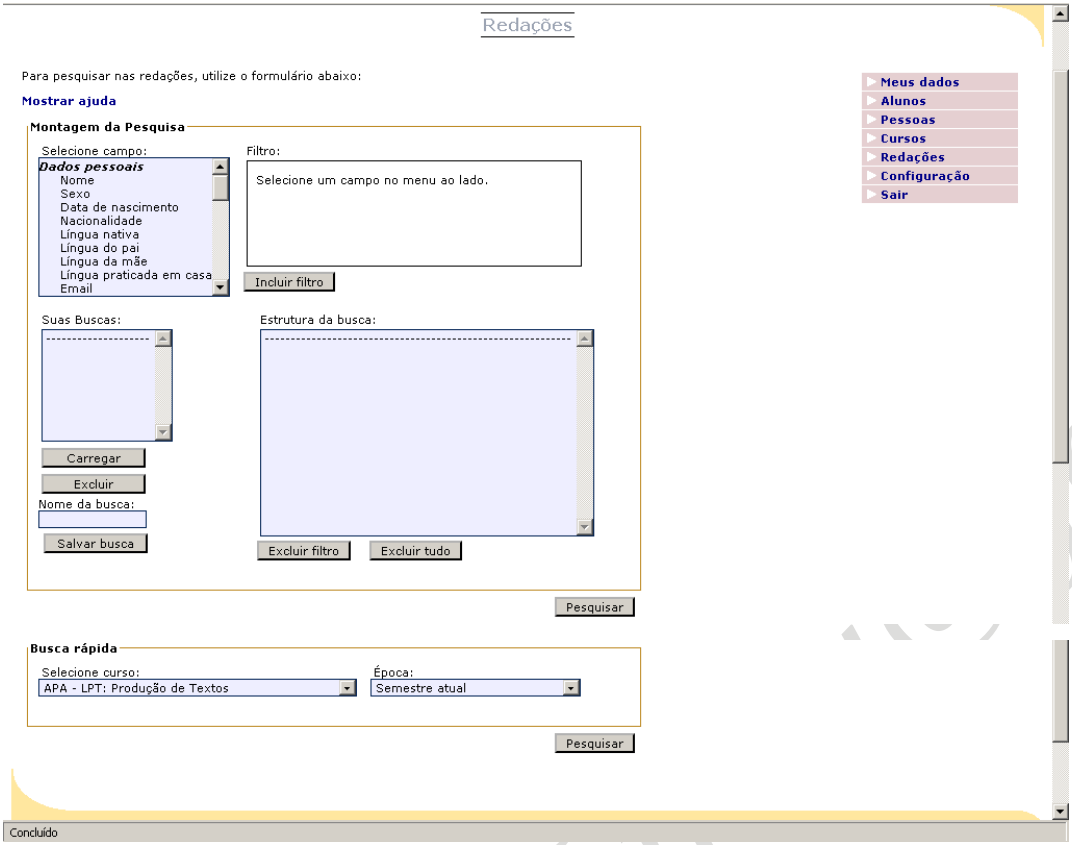

Figura 3 - Busca de redações através de filtros.

O pesquisador pode visualizar redação por redação (fig. 4) ou, se preferir, fazer o download do *corpus* de redações que solicitou (fig. 5).

Figura 4 - Leitura de redação na tela.

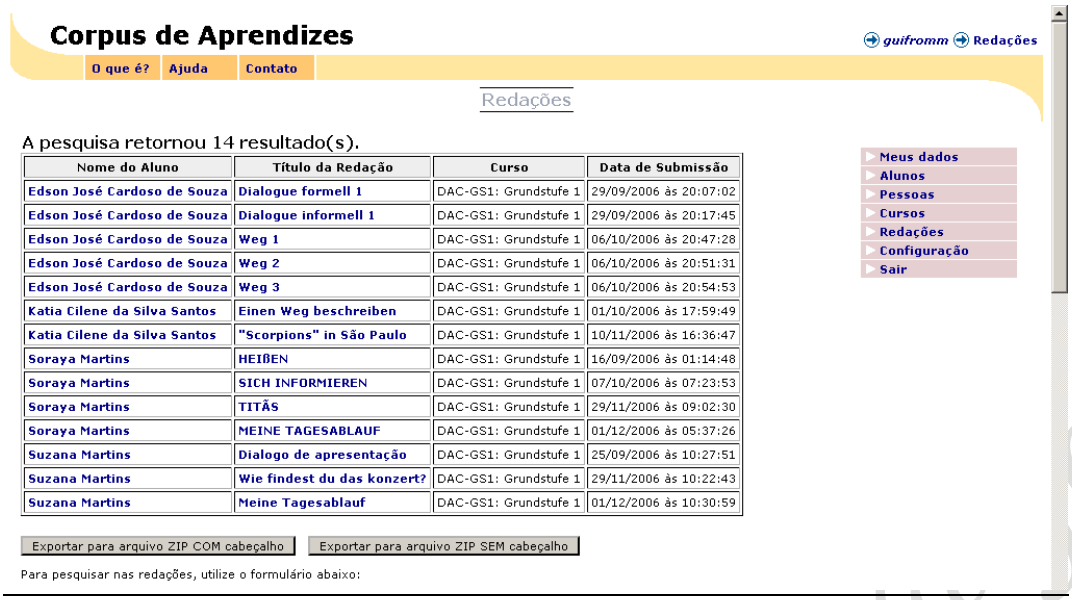

Figura 5 - *Corpus* de redações solicitadas e botão para exportação.

Essas redações, que podem ou não vir acompanhadas de cabeçalho, estão em formato .txt (para maior facilidade de trabalho com programas de análise lexical) e são compactadas, para envio, no formato .zip. Na figura 6, por exemplo, temos uma seleção de textos descompactados em um diretório cujo filtro de busca foi o sexo masculino. O código 29_EOC_B3_33.txt, por exemplo, representa o número de cadastro do aluno (29), o código do curso em questão (EOC – English On Campus), o estágio (B3 – Basic 3) e o número da redação (33 – número dentro do cômputo geral de redações cadastradas no sistema, em ordem de inclusão).

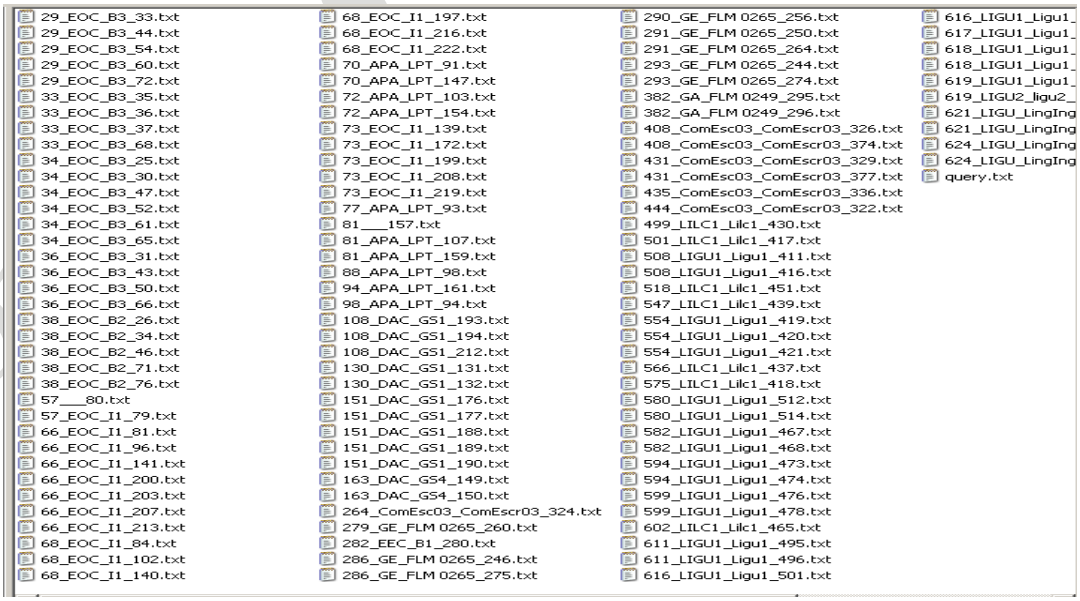

Figura 6 - redações exportadas.

A solicitação de cabeçalho gera arquivos que contêm, além da redação do aluno, todas as informações sobre o mesmo. O texto abaixo apresenta um exemplo (a redação contém apenas um excerto).

Redação extraída do *Corpus* de Aprendizes em 19/06/08 19:33:29

Aluno: XXXXXXXXXXXXX
Sexo: F
Data de Nascimento: 18/02/1959
Língua Materna: português
Faz outro curso de língua: S
Quais línguas: Espanhol
Anos de estudo: 0
Já morou no país da língua estrangeira que está estudando: S
Hábitos com a língua: internet,filmes,livros,revistas,outros
Escolaridade: Superior Incompleto
Curso: Pedagogia
Semestre atual: 8o. semestre
Instituição: Faculdade de Educação da USP (FE)
Ano de ingresso: 2003

Curso: EEC-B1: Básico 1
Semestre atual: 01/2007

Data de Submissão: 2007-06-12 15:29:50
Tipo: informativo
Assunto: Atrações turísticas
Referência: internet

Título: Buenos Aires

La ciudad de Buenos Aires es muy grande e encanta a los turistas por su arquitectura europea, su importante vida cultural con numerosos museos, salas de exposiciones y conferencias, galerías de arte, cines y teatros con espectáculos nacionales e internacionales de primer nivel.

Arquivo gerado em 19/06/08 19:33:29

Atualizações do Sistema

Foi solicitado (e aprovado) um novo financiamento ao CNPq (Processo 400988/2006-2) para correção e ampliação das funções do CoMAprend. Essas atualizações, já disponíveis, estão em fase de teste no site. Destacamos:

Inclusão de ferramentas de análise lexical

Foram integradas à busca três ferramentas que já estavam disponíveis para os usuários do CorTec[13]. A busca inicia-se a partir da seleção de textos, apertando-se o botão "Aplicar Ferramentas de Busca". Uma nova página se abre e mostra novamente todas as redações previamente selecionadas (Figura 7); é possível escolher uma, um grupo ou todas para fazer a análise. Após a seleção, basta clicar em "Próximo Passo" e escolher a ferramenta desejada (Figura 8):

a) Concordanciador: busca uma palavra, dentro do *corpus* selecionado, e cria linhas de concordância nas quais a palavra selecionada está centralizada;
b) gerador de lista de palavras: cria uma listagem com todas as palavras disponíveis no *corpus* criado com indicação de frequência;
c) gerador de n-gramas: sequência de palavras (de 1 a 4) a partir de uma dada palavra de entrada.

Figura 7. Primeira página para aplicação das ferramentas.

[13] *Corpus* Técnico. Apresenta *corpora* técnicos disponíveis em 14 áreas. Disponível para consulta em: www.fflch.usp.br/dlm/comet. A aplicação das ferramentas, inclusive, é realizada dentro do ambiente do CORTEC.

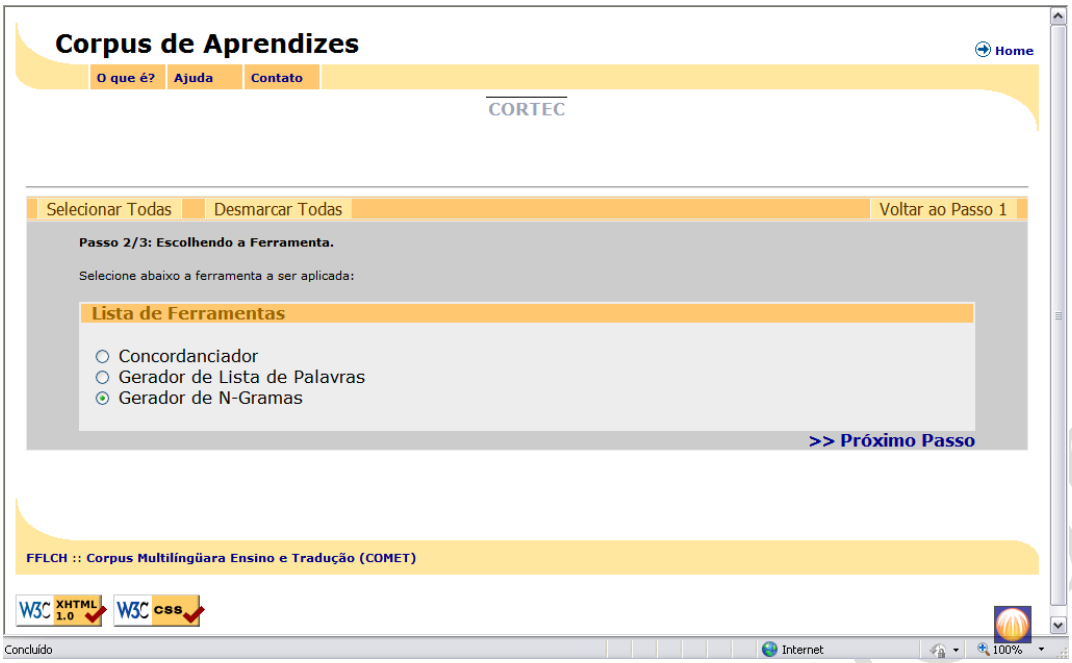

Figura 8. Ferramentas disponíveis – sites CoMAprend e CORTEC.

Ferramentas do administrador

Ferramentas e dados estatísticos foram adicionados à página do administrador. Entre essas alterações, destacamos:

a) Alteração de senha de pesquisadores e alunos. Ela é acessada somente pelo administrador e o objetivo é alterar as senhas já cadastradas, pois há constante perda de senha por parte dos usuários;

b) dados estatísticos gerais, para o administrador ter uma visão total do sistema;

c) inclusão/exclusão de novas instituições de ensino, línguas, nacionalidades.

Além da visualização de dados nesta página (fig. 9), foi incluída, também, a opção de o administrador visualizar, a partir do menu geral do lado direito, as instituições cadastradas no banco de dados.

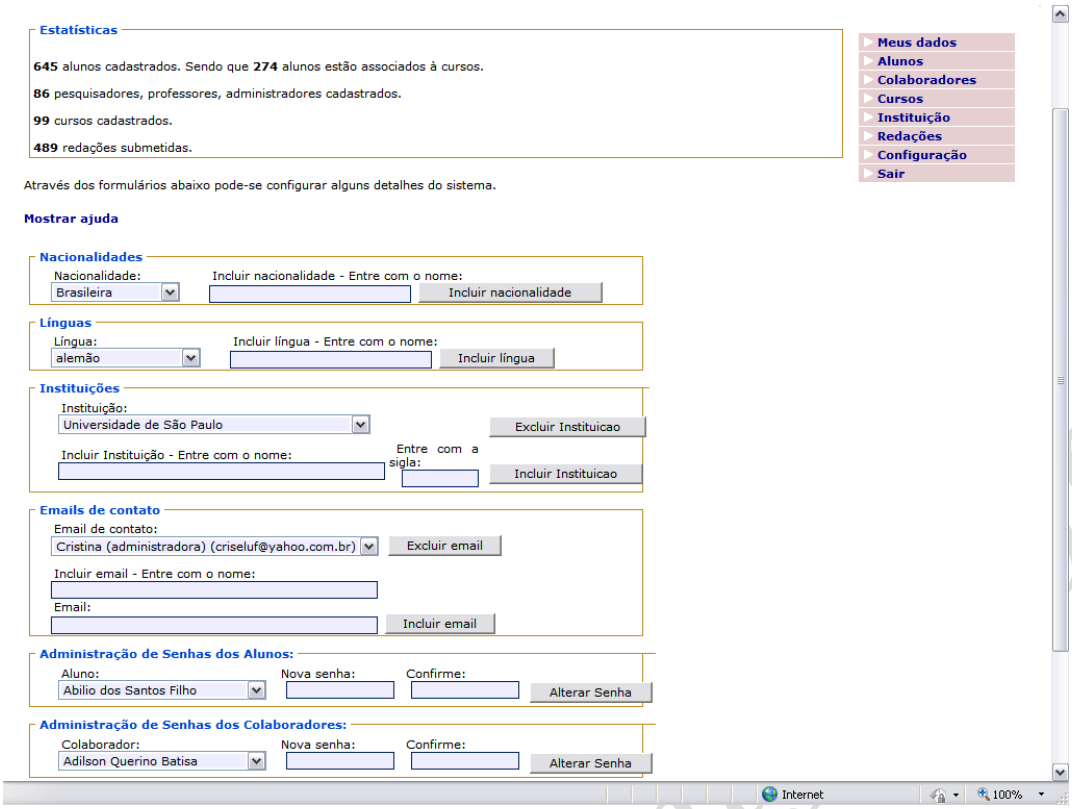

Figura 9. Configuração do sistema (somente para administradores).

Nova pesquisa

O *layout* da estrutura de pesquisas foi refeito, dividindo-a entre simples (fig.10), onde foi inserido o campo "língua" (pesquisa todas as redações escritas em determinada língua), e avançada (fig. 11).

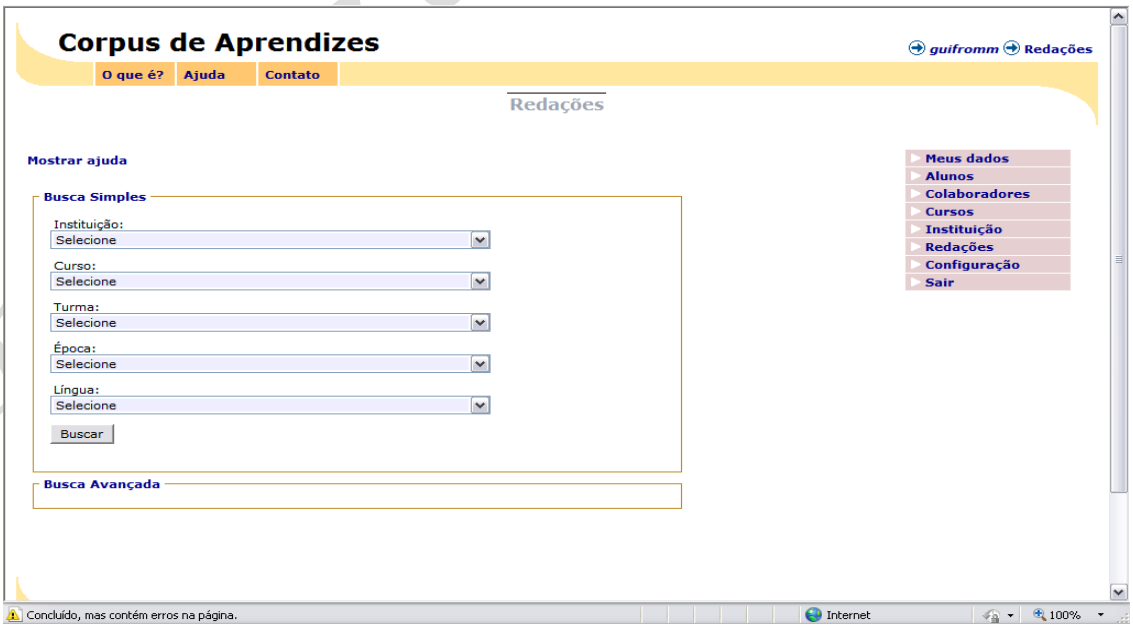

Figura 10. Tela de busca simples, novo *layout*.

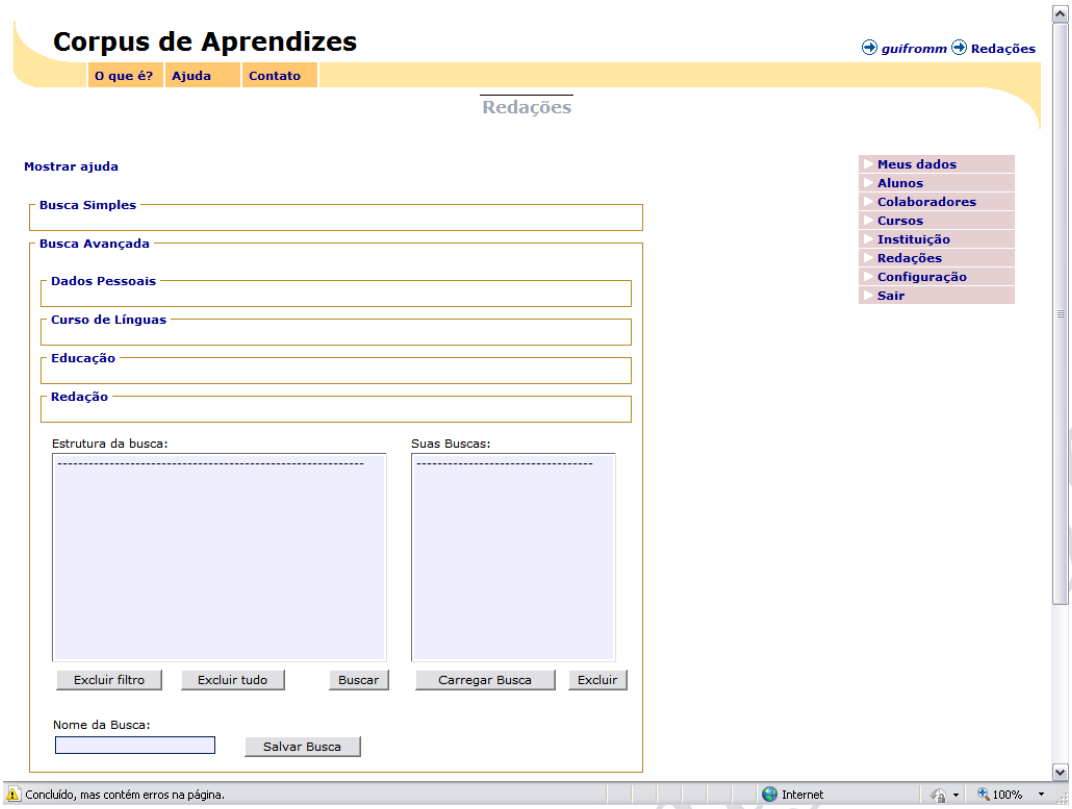

Fig. 11. Tela de busca avançada, novo layout.

Próximos Passos

A fase de testes da primeira versão gerou uma grande quantidade de cursos (e seus códigos) no sistema. Planeja-se, agora, antes da entrada do sistema em fase totalmente funcional, uma "limpeza" da base de dados. As redações já inseridas serão salvas[14], o banco de cursos do sistema será apagado e uma nova tabela de classificação de cursos, de acordo com cada instituição, deverá ser criada. Essa questão se mostrou preocupante, já que os professores enviam ao(s) administrador(es) diferentes nomes e códigos para os mesmos cursos. A padronização, neste nível, requer estreita cooperação entre o administrador e a nova instituição que estiver sendo cadastrada

Considerações Finais

O Corpus Multilíngue de Aprendizes, em testes na Universidade de São Paulo, não apenas integra as diferentes línguas ministradas pelo Departamento num projeto conjunto, como representa um projeto promissor em termos das possíveis áreas de pesquisas que enseja. De um projeto local, transformou-se, pela força da Internet, em

[14] As redações estão diretamente ligadas aos dados do curso. Se apagarmos o nome de um curso, por exemplo, todas as redações a ele associadas são apagadas também.

um projeto de amplo alcance. Qualquer professor ou pesquisador pode acessá-lo, associar-se a ele e gerenciar as redações inseridas por seus alunos no banco de dados do CoMAprend.

Referências Bibliográficas

ALTENBERG, B. **Advanced Swedish learners' use of causative *make*. A contrastive background study.** In Granger et al. 2002

ALTENBERG, B., TAPPER, M. *The use of adverbial connectors in advanced Swedish learners' written English*. In GRANGER, S. (ed.) **Learner English on Computer**. Addison Wesley Longman: London and New York, 1998. pp 80-93.

BERBER SARDINHA, T. **O Corpus de Aprendiz Br-Icle**, disponível em: http://lael.pucsp.br/~tony/2001bricle-interc.pdf, 2001.

____. **Lingüística de *corpus***. Barueri: Manole, 2004.

DE COCK, S. *A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-Native Speakers of English*. **International Journal of Corpus Linguistics** *vol 3(1)*: 1998. p. 59-80.

GILQUIN, G**.** *Linking up contrastive and learner corpus research: an introduction*. Paper presented at the workshop **Linking Up Contrastive and Learner Corpus Research**, University of Santiago de Compostela, 19 September 2005.

GILQUIN G., PAPP, S. & DÍEZ-BEDMAR, M.B. (eds). **Linking up Contrastive and Learner Corpus Research** . Amsterdam, Atlanta: Rodopi, 2008.

GRANGER, S. ***P**refabricated patterns in advanced EFL writing*: collocations and formulae. In COWIE, A. (ed.) **Phraseology**: theory, analysis and applications. Oxford: Oxford University Press, 1998a. pp 145-160.

____ **Learner English on Computer**. London & New York: Addison Wesley Longman, 1998b

____ *The Learner Corpus*: A Revolution in Applied Linguistics. **English Today** 39 (10/3), 1994a, p. 25-29.

____ *From CA to CIA and back*: An integrated approach to computerized bilingual and learner corpora. In Aijmer, Karin et alli (eds.) **Languages in Contrast – Papers from a Symposium on Text-based Cross-linguistic Studies**. Lund: Lund University Press, 1994b. p. 37-51.

___ *The International Corpus of Learner English*. In AARTS J., DE HAAN, P., OOSTDIJK, N. (eds) **English Language Corpora**: Design, Analysis and Exploitation. Amsterdam: Rodopi, 1993a, p. 57-69.

____ *The International Corpus of Learner English*. **The European English Messenger** 2(1), 34, 1993b.

GRANGER, S., F. MEUNIER (ed.). **Phraseology in Foreign Language Learning and Teaching**. Amsterdam: Benjamins, 2005.

KASZUBSKI, P. *Lexical profiling of English (learner) corpora*: can we measure advancement levels? In: LEWANDOWSKA-TOMASZCZYK, B., MELIA, P. (ed.) **Lódz studies in Language**, Vol. 1: **PALC'99***:* Practical Applications in Language Corpora [Papers from the International Conference at the University of Lódz, Poland, 15-18 April, 1999]. Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang, 2000. P. 249-86.

LEECH, G**.** *Learner corpora*: what they are and what can be done with them. In: GRANGER, S. (ed.) 1998b, p. xiv-xx.

TAGNIN, S. E. O. **O Jeito que a gente Diz**, São Paulo: Disal, 2005