O PROJETO COMET – CORPUS MULTILÍNGÜE PARA ENSINO E TRADUÇÃO – NA UNIVERSIDADE DE SÃO PAULO[1]

Stella E. O. TAGNIN (USP)[2]

RESUMO: Este artigo descreverá dois corpora do projeto COMET na USP: o CorTec e o CoMAprend. O **CorTec** é um corpus técnico bilíngüe inglês-português em cinco áreas – Culinária, Ecoturismo, Hipertensão, Informática e Instrumentos Contratuais –, cada uma com aproximadamente 200.000 palavras em cada língua, totalizando mais de 2 milhões de palavras. O site (http://www.fflch.usp.br/dlm/comet/consulta_cortec.html) produz concordâncias, listas de freqüência e n-gramas. O **CoMAprend** (Corpus Multilíngüe de Aprendizes) é alimentado com a produção de aprendizes de alemão, espanhol, francês, inglês e italiano dos cursos de graduação e extensão da USP (http://www.fflch.usp.br/dlm/comet/comaprend.html). No futuro terá ferramentas de busca semelhantes às do CorTec.

ABSTRACT: This paper will describe two corpora of the COMET Project at the University of São Paulo: CorTec and CoMAprend. **CorTec** is a technical English-Portuguese bilingual corpus in five domains: Computing, Commercial Contracts, Cooking, Ecotourism, and Hypertension, each containing at least 200,000 words in each language, which amounts to over 2 million words. The site (www.fflch.usp.br/dlm/comet) produces concordances, wordlists and n-grams. **CoMAprend** is a learner corpus which is being populated with the production of FL undergraduate and extension students of English, French, Italian, German, and Spanish. In due time it will also be searchable with computational tools.

1. Introdução

Os estudos com corpora vêm se expandindo no Brasil, mas mesmo assim ainda é pequena a produção acadêmica em relação à de outros países, em especial os europeus. Talvez um dos entraves seja a escassez, para não dizer praticamente a inexistência, de corpora especializados. Falta-nos, também, um corpus nacional, semelhante ao British National Corpus, ao American National Corpus ou ao Czech National Corpus, entre outros, cada um com, no mínimo, 100 milhões de palavras. Temos, é verdade, o Lácio-Ref, parte do projeto Lácio-Web (www.nilc.icmc.usp.br/lacioweb). Apesar de se tratar de um corpus de pequenas dimensões – aproximadamente 10 milhões de palavras –, se comparado aos corpora nacionais acima, tem a grande vantagem de estar totalmente disponível *on-line*, ou seja, o usuário tem acesso direto ao corpus por meio de concordâncias e listas de freqüência. O corpus também pode ser baixado na íntegra, ou em partes, para a máquina do usuário e então explorado com uma ferramenta específica, como o *software* WordSmith Tools (SCOTT 1996). Além de servir para um sem número de pesquisas lingüísticas, de lexicais a discursivas, presta-se também como corpus de referência de língua geral para os estudos terminológicos. É nestes últimos que se insere o **CorTec,** que passamos a descrever.

2. O CorTec

O **CorTec** (Corpus Técnico) é um dos corpora do projeto COMET (TAGNIN 2002a, 2002b, 2003a, 2003b), que está sendo desenvolvido na Universidade de São Paulo desde 2000. Abrange, por ora, cinco áreas – Culinária, Ecoturismo, Hipertensão, Informática e Instrumentos Contratuais – e contém textos originais em inglês e português, ou seja, trata-se de um corpus comparável (ULRYCH 1997), pois o material é compilado seguindo critérios semelhantes quanto ao gênero, ao conteúdo, à extensão, à função comunicativa, entre outros. Cada corpus contém, no mínimo, 200.000 palavras, em cada língua, o que totaliza mais de 2 milhões de palavras.

[1] Meus agradecimentos ao CNPq, que viabilizou a disponibilização do CorTec (Processo 403120/2003-9) e está financiando a implementação das ferramentas do CoMAprend (CNPq Processo 400988/2006-2).
[2] E-mail para contato: seotagni@usp.br.

2.1 O Corpus de Culinária

Esse corpus é formado essencialmente por receitas coletadas da Internet, 1.555 receitas em português brasileiro e 2.076 em inglês britânico (TEIXEIRA 2005). Embora tenha sido feito um esforço para garantir que as receitas tenham sido escritas originalmente nessas línguas, é possível que uma ou outra seja traduzida, pois, com a globalização, isso se tornou comum nessa área. Outro critério que orientou a coleta desse material foi assegurar que todas as categorias culinárias (aperitivos, acompanhamentos, pratos de massa, de peixe, de carne, sobremesas etc.) estivessem representadas. A parte em inglês tem 368.227 ocorrências e a em português, 252.149. O type/token ratio em ambas as línguas – 1,98 em inglês e 2,84 em português – aponta para uma linguagem bastante repetitiva.

2.2 O Corpus de Ecoturismo

Esse corpus é composto por textos retirados de sites do governo, de entidades ambientalistas e de agências de turismo, em especial do Brasil, dos Estados Unidos e da Grã-Bretanha (MARTINS 2005). Contém 201.826 palavras em inglês e 200.887 em português. É interessante notar que o type/token ratio em inglês é 4,96, enquanto em português é quase o dobro, 8,93, o que indica maior riqueza lexical – menos repetições – na nossa língua.

2.3 O Corpus de Hipertensão

Esse corpus é constituído de artigos coletados de periódicos e revistas brasileiros e americanos, sendo126 textos em inglês e 125 em português (CASTANHO & GINEZI). Devido à abrangência da **hipertensão**, os textos em que ela é discutida podem pertencer a áreas como a Cardiologia, a Saúde Pública, a Nefrologia, entre outras. O corpus em inglês contém 453.475 palavras e o em português 356.718, com um type/token ratio de 3,93 e 6,17, respectivamente. Novamente, os textos em português apresentam menos repetições.

2.4 O Corpus de Informática

O Corpus de **Informática** (FROMM 2005) foi compilado exclusivamente com textos de publicações na Internet que contemplam a área geral de Tecnologia de Informação (TI). Contém 193.877 palavras em inglês e 196.604 em português, com um type/token ratio de 6,66 e 7,72, respectivamente. Ao contrário dos outros corpora, a diferença entre as duas línguas nesta área, em termos de repetições, parece ser mínima.

2.5 O Corpus de Instrumentos Contratuais

Esse corpus é composto de 48 Instrumentos Contratuais em inglês e 134 em português (CARVALHO 2005). O conceito do que é *contrato* no direito brasileiro orientou a seleção dos documentos, de modo que foram incluídos os seguintes tipos de documentos: Contratos de Compra e Venda, Contratos de Prestação de Serviços, Contratos de Distribuição, Contratos de Locação, Contratos de Licença, Contratos de Fornecimento, Contratos Sociais, Contratos Bancários, Contratos de Empréstimo, Contratos de Franquia, Procurações, Pactos Antenupciais e Termos de Sigilo. Apesar da discrepância entre a quantidade de documentos nas duas línguas, o número total de palavras nos dois corpora é bastante próximo: 204.249 em inglês e 200.588 em português, o que parece indicar que os documentos em inglês tendem a ser bem mais extensos. Já o type/token ratio, 2.96 em inglês e 4.83 em português aponta novamente para uma maior diversidade lexical nesta última.

2.6 O acesso aos corpora

Os corpora podem ser estudados por meio de três ferramentas: Concordanciador, Gerador de Lista de Palavras, Gerador de N-Gramas.

2.6.1 Concordanciador

Por questões de direitos autorais, o usuário não tem acesso aos textos na íntegra, mas pode consultar os corpora, individualmente ou em combinação, por meio de concordâncias, que permitem buscas por

palavras específicas (Igual a), por prefixos ou início de palavra (Começando com), por sufixos ou finais de palavra (Terminando com) ou ainda por trechos internos de palavras (Contendo) (vide Fig 1):

Fig. 1: CorTec – Parte da tela de seleção da expressão de busca

As linhas de concordância apresentam a expressão de busca centralizada e um co-texto de até 60 caracteres de cada lado.

Foram encontradas 1370 ocorrências!! Fazer DOWNLOAD do resultado.

Clique na palavra de busca para obter um contexto expandido com 150 caracteres

1 o reflete a dos músculos esqueléticos e cardiacos 15, 18, 20 , foi propósito deste tr
2 de pressão 7 . A administração de Mg em cardiopatas causa a redução da resistência va
3 s e considerados desprovidos de efeito "cardioprotetor". Entretanto, também é verdade
4 nérgicos, repercutem sobre a fisiologia cardiovascular e exigem especial consideração
5 e na redução da morbidade e mortalidade cardiovascular. Os efeitos colaterais como
6 padores de potássio é indicada para os cardiopatas, especialmente quando medicados c
7 razão hemodinâmica e metabólica 21 . Os cardioseletivos causam menor fadiga e menor e
8 mg ao dia, e do metoprolol iedurilesl) (cardioseletivo), 200 mg ao dia, foi bem toler
9 opatia, de- vido ao seu possível efeito cardioprotetor. Podem ser úteis, ainda, em po
10 r de risco para morbidade e mortalidade cardiovasculares 28 (insuficiência cardíaca e
11SO AMODEO Instituto Dante Pazzanese de CardiologiaŠSP Correspondência: Dr. Celso A

Fig. 2: CorTec – Linhas de concordância para “cardi” no Corpus de Hipertensão

Ao clicar na expressão de busca, o formato é ampliado e o co-texto passa para cerca de 150 caracteres de cada lado.

Ocorrência 1
s mono-nucleares
(linfocitário) e pela retenção do Mg infundido ló,
17 . Como a concentração de Mg linfocitário reflete a dos
músculos esqueléticos e cardiacos 15, 18, 20 , foi propósito
deste trabalho o estudo do Mg sérico e linfocitário em idosos
sadios e hipertensos.
CASUÍSTICA E MÉTODOS
Foram selecio

Fig. 3: CorTec – Linha de concordância ampliada

Como se observa na Fig. 2, o programa fornece o número total de ocorrências – 1370 nesse caso – e a possibilidade de fazer o *download* do resultado total, o que permite que o usuário possa manipular esse material com um programa de investigação de corpus, como o WordSmith Tools. Isso lhe permitiria ordenar

as linhas pelo contexto à direita ou à esquerda da palavra de busca, facilitando a visualização de eventuais padrões lexicais ou sintáticos, além de utilizar as outras ferramentas disponíveis, como *Collocates* e *Clusters*.

2.6.2 Gerador de Lista de Palavras

Outra ferramenta embutida no CorTec é o Gerador de Lista de Palavras, que fornece dois tipos de lista, uma por ordem de freqüência, outra em ordem alfabética. Fornece também um resumo do léxico, considerando a lexia complexa de nomes próprios, ou seja, considera Universidade de São Paulo uma só palavra, por exemplo. A exemplo das linhas de concordância, a lista de palavras também pode ser baixada para a máquina do usuário.

Fazer DOWNLOAD do Resultado.

Ocorrências (tokens), considerando a Lexia Complexa de Nomes Próprios
Total de Ocorrências: 442275
Total de Ocorrências que aparecem uma vez: 19133
Total de Ocorrências que aparecem mais de uma vez: 423142

Palavras únicas/formas (types), considerando a Lexia Complexa de Nomes Próprios
Total de Palavras: 35058
Total de Palavras que aparecem uma vez: 17259
Total de Palavras que aparecem mais de uma vez: 17799

Fig. 4: CorTec – Resumo das ocorrências do Corpus de Hipertensão

Como em toda lista de palavras, as de maior freqüência são as palavras gramaticais (eliminadas da listagem abaixo). No caso em pauta, a primeira palavra de conteúdo é “pacientes”, na 25ª. posição, com 1663 ocorrências. Seguem-se “arterial” na 35ª, com 1289 ocorrências, “risco” na 47ª, com 840 e “hipertensão” na 49ª, com 826 ocorrências.

25-pacientes	1663
26-como	1588
27-um	1563
28-is	1502
29-mais	1470
30-with[3]	1459
31-à	1440
32-se	1371
33-ao	1305
34-foi	1301
35-arterial	1289
36-for	1275
37-uma	1226
38-ser	1153
39-nos	1151
40-das	1085
41-that	1037
42-entre	1036
43-foram	1035
44-pressure	978
45-blood	955
46-são	942
47-risco	840
48-or	827
49-hipertensão	826
50-estudo	803

[3] As palavras em inglês que aparecem indevidamente na lista acima podem fazer parte do *abstract* dos artigos, de referências bibliográficas ou ainda de citações dentro do texto, e devem ser ignoradas num levantamento do vocabulário em português. A bem da verdade, esse problema deve ser sanado na próxima etapa do CorTec, que justamente prevê a correção de falhas, a ampliação dos corpora existentes, assim como a inclusão de novos corpora.

Tabela 1: Parte da Lista de Palavras por ordem de freqüência do corpus de hipertensão em português

Como se trata de um corpus bilíngüe, pode-se gerar uma lista de freqüência também em inglês para comparar as palavras de maior ocorrência.

19-patients	1944
20-this	1760
21-be	1736
22-not	1707
23-pressure	1616
24-blood	1603
25-are	1518
26-an	1397
27-study	1328
28-hypertension	1270

Tabela 2: Parte da Lista de Palavras por ordem de freqüência do corpus de hipertensão em inglês

O que se observa é que a primeira palavra de conteúdo, *patients*, é a mesma que em português, mas já ocorre na 19ª posição, com 1944 ocorrências. Em seguida ocorrem *pressure* (23ª posição) e *blood* (24ª posição), com praticamente a mesma freqüência, o que poderia indicar alta probabilidade de co-ocorrência: *blood pressure*. Antes de *hypertension* (28ª posição), ainda ocorre *study* (27ª posição), apontando, provavelmente para um elevado número de textos acadêmicos.

Se recorrermos às concordâncias para nos certificarmos da alta freqüência da colocação *blood pressure*, seremos surpreendidos por apenas 146 ocorrências! Uma análise mais minuciosa revelará a co-ocorrência de *pressure* com *venous* (20 oc.), *systolic* (34 oc.) e *diastolic* (40 oc.), entre outras. Já *blood* ocorre com certa freqüência com *flow* (55 oc.), e em número bem menor com *count* (4 oc.), *sugar* (6 oc.), *cholesterol* (7 oc.) e *vessel* (8 oc.). Observa-se, assim, que sempre convém verificar uma hipótese sugerida pela Lista de Freqüência fazendo uso das linhas de concordância, para certificar-se de que ela se confirma (ou não, como neste caso).

2.6.3 Gerador de N-Gramas

N-Gramas são seqüências de cadeias de palavras, de modo que um bigrama é uma seqüência de duas cadeias de caracteres, um trigrama, de três e assim por diante. O CorTec permite a extração de n-gramas com 2, 3 ou 4 cadeias.

Abaixo, os primeiros 26 bigramas do corpus de hipertensão em inglês, com freqüência mínima de 3 ocorrências:

1. blood pressure	2. Ang II
3. mm Hg	4. big ET-
5. present study	6. left ventricular
7. mmol L	8. hypertensive patients
9. In addition	10. mol L
11. mg kg	12. coronary artery
13. heart failure	14. de palavras
15. Ang I	16. Número de
17. arterial pressure	18. heart rate
19. essential hypertension	20. hypertensive subjects
21. angiotensin II	22. body weight
23. diastolic blood	24. Blood Pressure
25. risk factors	26. control subjects

Tabela 3: Os primeiros 26 bigramas do corpus de hipertensão em inglês

Nota-se que *blood pressure* é o bigrama mais freqüente, mas que o mesmo aparece novamente na posição 24, grafado com letra maiúscula, porque o gerador faz uma diferença entre letras maiúsculas e minúsculas. Observa-se, também, *de palavras* (14ª posição) e *Número de* (16ª posição), indicando "sujeira" no corpus, ou seja, não foram eliminados trechos em português dentro do corpus em inglês.

É preciso atentar para o fato de que nem todo bigrama é uma unidade lexical, cabendo ao pesquisador fazer essa seleção. Isso fica ainda mais patente numa listagem de trigramas, como se pode verificar abaixo:

1. in patients with	2. the present study
3. in blood pressure	4. Número de palavras
5. the effects of	6. of Ang II
7. the presence of	8. weeks of age
9. the development of	10. blood pressure and
11. and Ang II	12. of big ET-
13. in response to	14. men and women
15. American Heart Association	16. the effect of
17. blood pressure in	18. systolic and diastolic
19. in the present	20. been shown to
21. diastolic blood pressure	22. systolic blood pressure
23. sympathetic nerve activity	24. of patients with
25. the increase in	26. to Ang II

Tabela 4: Os primeiros 26 trigramas do corpus de hipertensão em inglês

Ao que parece, há apenas 2 trigramas que formam uma unidade lexical: *American Heart Association* e *diastolic blood pressure*; se bem que podem, eventualmente, fazer parte de uma unidade maior. Ao fazer uma concordância para *diastolic blood pressure*, por exemplo, encontramos 4 ocorrências de *supine diastolic blood pressure*, o que parece indicar uma provável nova unidade lexical.

Como se observa, um dado encontrado pode levar a uma investigação mais aprofundada, a um ir-e-vir entre as ferramentas em busca de mais informações, mais detalhes. Esse processo é chamado de pesquisa direcionada pelo corpus, em oposição a pesquisa baseada em corpus, na qual busca-se apenas a confirmação – ou não – de uma hipótese previamente estabelecida, a exemplificação para uma palavra ou expressão, sem enveredar por novos caminhos que porventura os dados venham a apontar.

2.7 Possibilidades de pesquisa

O CorTec presta-se, em especial, a pesquisas lexicais e discursivas, que incluem desde a extração de exemplos para os itens lexicais que se deseja pesquisar, busca de palavras ou grupos de palavras mais recorrentes em determinado corpus, até a pesquisa de tempos e formas verbais que caracterizam, por

exemplo, o discurso de uma das áreas de especialidade representadas. (LARANJINHA 1999, Bowker 1999, 2002, Bowker & Pearson 2002, Tagnin 2002c, 2002d).

3. O CoMAprend

Outra área de pesquisa em que o recurso a corpora oferece grande possibilidade de pesquisa é a do ensino e aprendizagem de línguas, quer maternas quer estrangeiras. Nesse âmbito, Tim Johns (1991a, 1991b) introduziu o conceito de DDL (*Data Driven Learning* - aprendizado por meio de dados) método que se vale de concordâncias produzidas a partir de um corpus de língua geral. Debruçando-se sobre essas concordâncias – em diversos formatos (http://www.eisu.bham.ac.uk/johnstf/unl_ddl.htm) – o aluno faz inferências a respeito, por exemplo, de determinadas colocações lexicais ou padrões sintáticos em que certo item lexical ocorre. Essa metodologia privilegia uma abordagem cognitiva do aprendizado, permitindo que o aluno aprenda "por descoberta" e não por memorização.

Ainda no âmbito do ensino, os corpora de aprendizes (GRANGER 2002) têm-se mostrado de extrema importância para detectar áreas de dificuldade dos alunos, seus erros mais freqüentes, permitindo, dessa forma, a elaboração de material didático específico para se obter um aprendizado mais eficaz.

O CoMAprend, que está sendo compilado junto ao Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, reúne a produção dos alunos de graduação e de extensão, em especial dos cursos de língua estrangeira "no campus". Esses cursos oferecem o aprendizado das cinco línguas do Departamento – alemão, espanhol, francês, inglês e italiano – para alunos de diversos níveis, desde principiantes até avançados.

O material está sendo coletado no próprio site do corpus, que pode ser acessado em http://www.fflch.usp.br/dlm/comet/comaprend.html. Inicialmente, o professor deve solicitar seu cadastro ao administrador do sistema, via e-mail ou pessoalmente. Em seguida, o professor fornece-lhe uma listagem com seus alunos e as disciplinas que está ministrando. Feito isso, o administrador cadastra a disciplina e associa os respectivos alunos a ela. A inscrição do aluno, feita diretamente no site, pressupõe o preenchimento de uma ficha cadastral que já inclui uma autorização para que o material inserido no corpus possa ser usado para fins de pesquisa. As informações dos alunos servirão como parâmetros para o recorte do corpus que o professor queira pesquisar. Concluídos os cadastros, o aluno pode acessar ao site, que tem um espaço específico para inserção de suas redações.

O **CoMAprend** permitirá a realização de um grande número de pesquisas, sobretudo nos campos do ensino, aprendizagem e aquisição da língua estrangeira. No âmbito da aprendizagem ou aquisição da língua estrangeira, poderão ser investigados aspectos morfológicos, sintáticos, lexicais e, até mesmo, pragmáticos e discursivos da linguagem dos aprendizes (sua interlíngua), por meio da análise da sua produção escrita. Atualmente, as disciplinas da graduação que abordam aspectos sintáticos, semânticos ou discursivos das línguas estrangeiras carecem de um banco de textos desse tipo, que possibilite a consecução de pesquisas aplicadas com base em dados autênticos de uso da língua estrangeira por aprendizes, já sistematizados e facilmente acessíveis por meio de ferramentas de busca.

Já para os docentes e alunos de pós-graduação (mestrado e doutorado), além das investigações referentes a aspectos da aprendizagem ou aquisição da língua, o **CoMAprend** possibilitará também o desenvolvimento de pesquisas relativas ao ensino da língua estrangeira, no que diz respeito à detecção e análise de erros dos aprendizes ou das suas dificuldades mais freqüentes, com vistas à elaboração de materiais didáticos específicos ou à análise de materiais já existentes. Tanto no âmbito da aprendizagem/aquisição, quanto no do ensino da língua estrangeira, esses estudos produzirão um conhecimento específico sobre os aprendizes brasileiros, com base em dados reais da sua produção lingüística.

Uma vez que o corpus será constituído, em parte, pelas redações dos alunos dos cursos de extensão universitária *no Campus*, essa fonte de informações estará disponível também para pesquisas que os professores ministrantes desses cursos desejem fazer com propósitos mais práticos e imediatos de levantamento das áreas de dificuldades dos alunos para a elaboração de programas de curso mais eficazes, assim como para a análise e preparação de materiais didáticos.

3.1 O acesso ao corpus

Somente os professores ou pesquisadores ligados ao projeto poderão acessar a produção dos alunos. Após fazer o *login*, surgirá uma tela que lhes permitirá estabelecer o recorte de sua pesquisa, ou seja, selecionar as variáveis que orientarão a construção de seu corpus, tais como: nome, sexo, data de

nascimento, nacionalidade, língua nativa, língua do pai, língua da mãe e língua praticada em casa. No exemplo abaixo, foram selecionados: língua nativa = português, sexo = feminino. Nesse caso, então, o corpus de estudo será constituído da produção dos alunos do sexo feminino cuja língua nativa é o português (vide Fig. 5). O passo seguinte é a seleção das turmas (vide Fig. 6). Estabelecido o perfil do aluno, a busca retornará, nas turmas selecionadas, as redações dos alunos com esse perfil (vide Fig. 7).

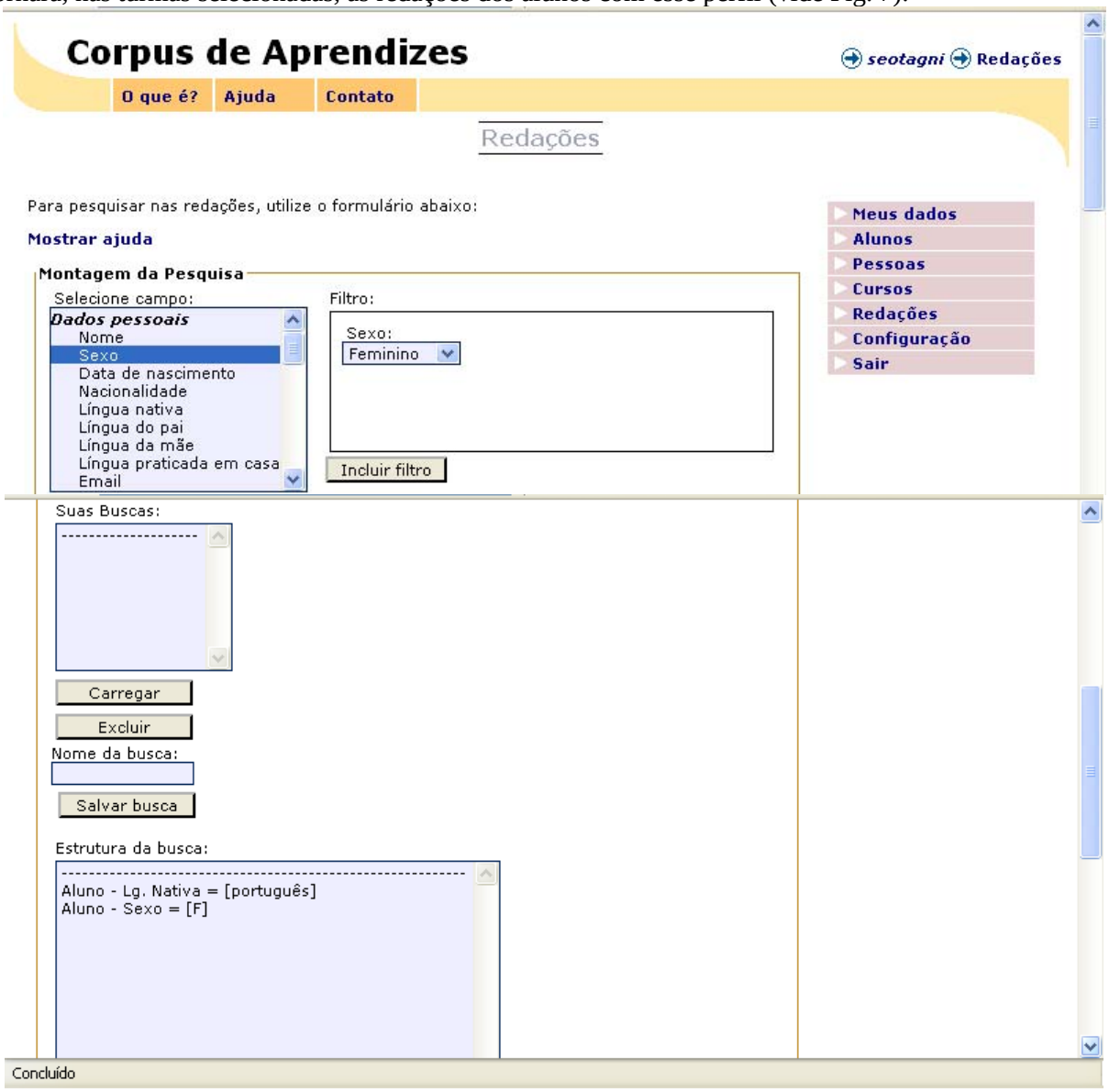


Fig. 5: CoMAprend – Parte da tela com opções para construção do corpus de pesquisa.

Fig. 6: CoMAprend – Parte da tela com opções para seleção das turmas a integrarem o corpus de pesquisa.

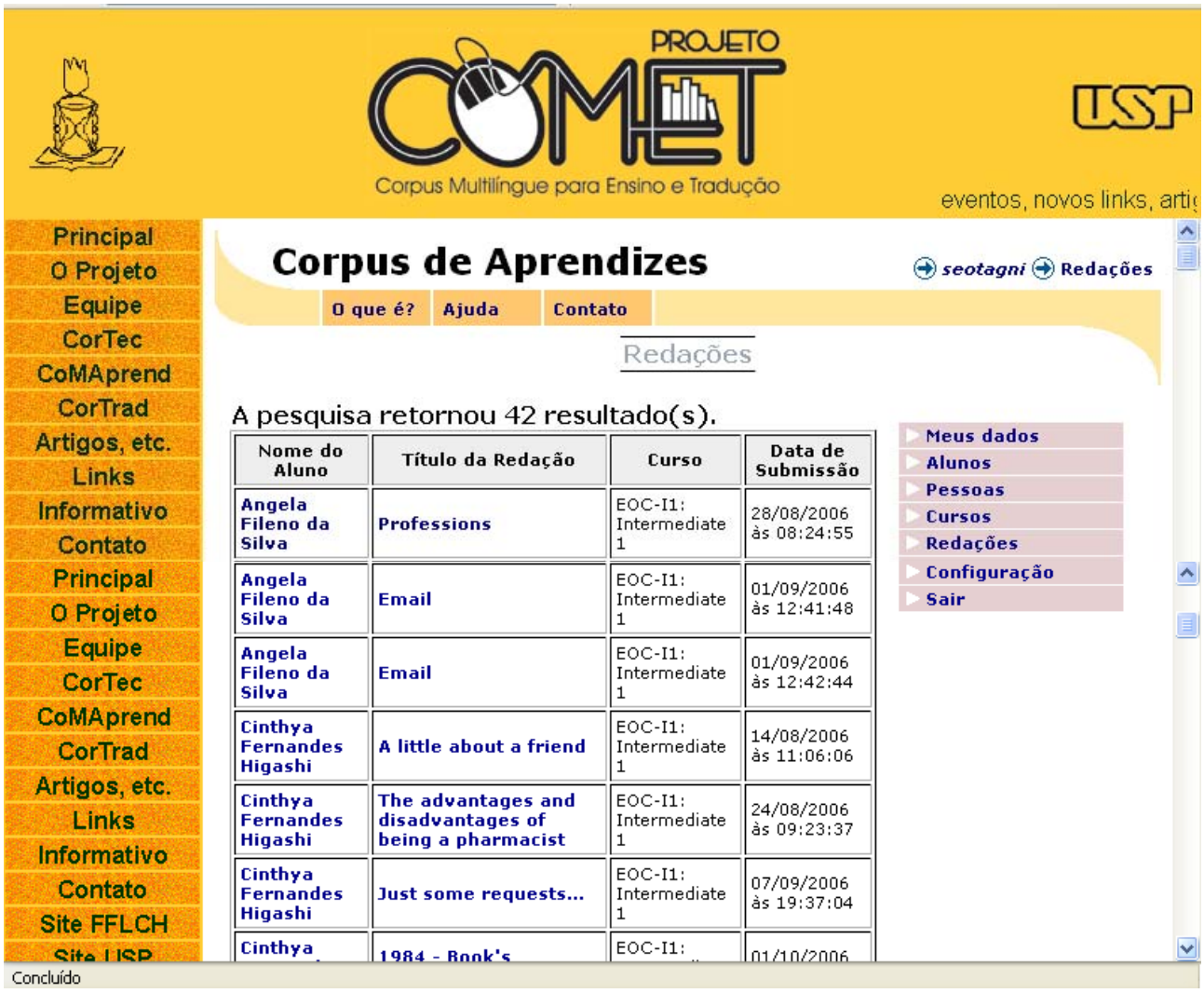
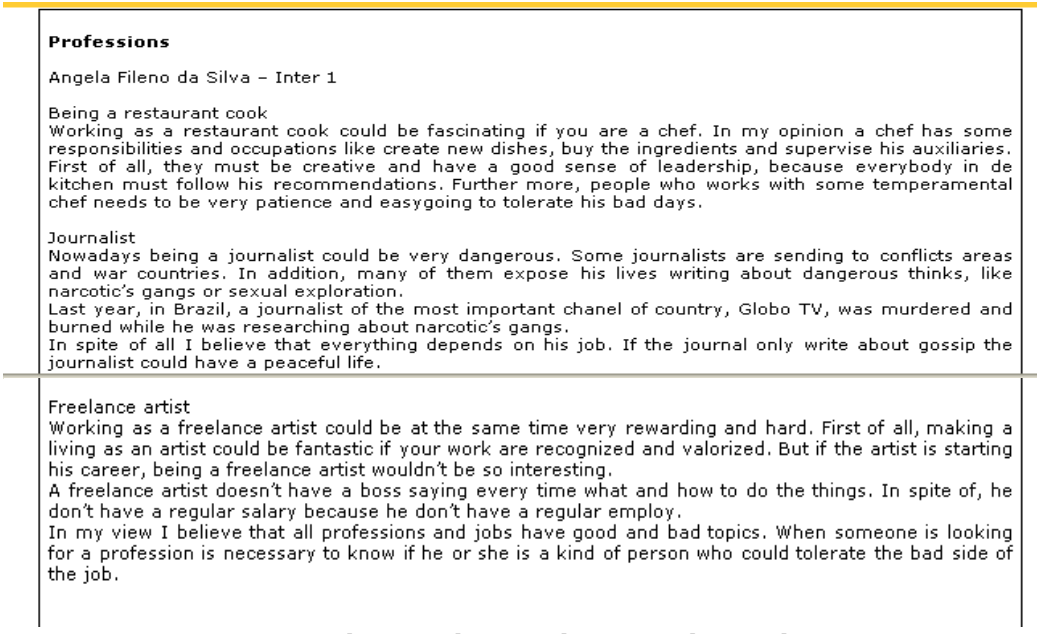


Fig. 7: CoMAprend – Tela com os resultados da pesquisa.

Basta clicar sobre o título da redação e ela aparecerá na tela:



Professions

Angela Fileno da Silva – Inter 1

Being a restaurant cook
Working as a restaurant cook could be fascinating if you are a chef. In my opinion a chef has some responsibilities and occupations like create new dishes, buy the ingredients and supervise his auxiliaries. First of all, they must be creative and have a good sense of leadership, because everybody in de kitchen must follow his recommendations. Further more, people who works with some temperamental chef needs to be very patience and easygoing to tolerate his bad days.

Journalist
Nowadays being a journalist could be very dangerous. Some journalists are sending to conflicts areas and war countries. In addition, many of them expose his lives writing about dangerous thinks, like narcotic's gangs or sexual exploration.
Last year, in Brazil, a journalist of the most important chanel of country, Globo TV, was murdered and burned while he was researching about narcotic's gangs.
In spite of all I believe that everything depends on his job. If the journal only write about gossip the journalist could have a peaceful life.

Freelance artist
Working as a freelance artist could be at the same time very rewarding and hard. First of all, making a living as an artist could be fantastic if your work are recognized and valorized. But if the artist is starting his career, being a freelance artist wouldn't be so interesting.
A freelance artist doesn't have a boss saying every time what and how to do the things. In spite of, he don't have a regular salary because he don't have a regular employ.
In my view I believe that all professions and jobs have good and bad topics. When someone is looking for a profession is necessary to know if he or she is a kind of person who could tolerate the bad side of the job.

Fig. 8. CoMAprend – Uma das 42 redações resultantes da pesquisa.

A seguir o pesquisador pode salvar as redações em formato *.txt* para constituir o seu corpus de pesquisa, que poderá ser então investigado com ferramentas computacionais específicas como é o caso do WordSmith Tools (SCOTT 1996). Num futuro próximo, o CoMAprend também contará com ferramentas semelhantes às do CorTec, o que permitirá algumas pesquisas *on-line*.

4. Considerações finais

Este artigo apresentou dois corpora que fazem parte do Projeto COMET, em desenvolvimento na Universidade de São Paulo. O primeiro é o CorTec, um corpus técnico bilíngüe, inglês-português, que abrange cinco áreas de especialidade: Culinária, Ecoturismo, Hipertensão, Informática e Instrumentos Contratuais. O site permite a produção de listas de freqüência, linhas de concordância e listas de n-gramas. Destina-se a pesquisas lexicais e discursivas. O segundo é o CoMAprend, um corpus de aprendizes de cinco línguas estrangeiras: alemão, espanhol, francês, inglês e italiano. Contém redações tanto de alunos da graduação quanto dos cursos de línguas extracurriculares. Permite identificar áreas de dificuldade dos aprendizes, seus erros mais comuns – dados que poderão informar a produção de material didático mais adequado a esses aprendizes.

5. Referências Bibliográficas

BOWKER, Lynn & PEARSON, Jeniffer. **Working with Specialized Language – A practical guide to using corpora**, London/New York: Routledge, 2002.

BOWKER, Lynn. Exploring the Potential of Corpora for Raising Language Awareness in Student Translators. **Language Awareness** vol 8, no. 3&4, 160-173, 1999.

BOWKER, Lynn. **Computer-Aided Translation Technology: A Practical Introduction**. Ottawa: University of Ottawa Press, 2002.

CARVALHO, Luciana. Compiladora do corpus de Instrumentos Contratuais/CorTec: www.fflch.usp.br/dlm/comet, 2005.

CASTANHO, Rosa Caporrino & GINEZI, Luciana Compiladoras do corpus de Hipertensão Arterial/CorTec: www.fflch.usp.br/dlm/comet, 2005.

FROMM, Guilherme. Compilador do corpus de Informática/CorTec: www.fflch.usp.br/dlm/comet, 2005.

GRANGER, S. A Bird's-eye View of Computer Learner Corpus Research, in: GRANGER, S. et al (Eds.) **Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching**. Amsterdam and Philadelphia: Benjamins. 3-33**,** 2002.

JOHNS, Tim. From printout to handout: grammar and vocabulary teaching in the context of data-driven learning**. ELR Journal (New Series) vol 4, Classroom Concordancing**, The University of Birmingham, 27-45, 1991a.

JOHNS, Tim. Should you be persuaded – two samples of data-driven learning materials. **ELR Journal (New Series) vol 4, Classroom Concordancing**, The University of Birmingham, 1-16, 1991b.

LARANJINHA, Ana Lucinda Tadei. **Para um Glossário Bilíngüe – Português/Inglês de Termos do Direito Comercial: Colocações Verbais**, dissertação de mestrado, Universidade de São Paulo, 1999.

MARTINS, Josimeire. Compiladora do corpus de Ecoturismo/CorTec: www.fflch.usp.br/dlm/comet. 2005.

SCOTT, M. **Wordsmith Tools**, Oxford: Oxford University Press, 1996.

TAGNIN, S.E.O. COMET - um corpus multilíngüe para ensino e tradução**.** In **Boletim da ABRALIN, vol. 2,** março de 2001, Anais do II Congresso Internacional da ABRALIN, Fortaleza, 13 a 16 de março de 2001, 589-591, 2003a.

TAGNIN, S. E. O. COMET – A Multilingual Corpus for Teaching and Translation. In: Lewandowska-Tomaszczyk (ed**.) PALC 2001 – Practical Applications in Language Corpora** (Proceedings of the International Conference on Practical Applications in Language Corpora, Lodz, Poland. September 07-09, 2001). Frankfurt-am-Main:Peter Lang, 535-540, 2003b.

TAGNIN, S. E. O.. A multilingual learner corpus in Brazil, In Archer, Dawn, Paul Rayson, Andrew Wilson and Tony McEnery (eds.), **Proceedings of the Corpus Linguistics 2003 UCREL Technical Papers Vol. 16, Part 1,** Special Issue (2003); ISBN 1 86220 131 5, Lancaster University (UK) 28-31 March 2003, 940-945, 2003c

TAGNIN, S.E.O. Um corpus de integração: o projeto COMET, comunicação apresentada no **5o. Simpósio de Lingüística de Corpus, 2002, Criação, anotação e aplicação de corpora** (InPLA, PUC/SP), São Paulo, 25-26/04/2002, 2002a.

TAGNIN, S.E.O. Taking off in Brazil: COMET – A Multilingual Corpus for Teaching and Translation. In Aijmer, Karin (ed.) **ICAME 2002: The Theory and Use of Corpora** - The 23rd International Conference on English Language Research on Computerized Corpora of Modern and Medieval English – Göteborg, 22-26 May 2002, Gotemburgo, Suécia, 67, 2002b.

TAGNIN, S.E.O.. Corpora and the Innocent Translator: How can they help him, in THIELEN, Marcel (Ed*.)* **Translation and Meaning, Part 6, Proceedings of the Lodz Session of the 3rd Maastricht-Lodz Duo Colloquium on "Translation on Meaning"**, Lodz, Poland, September 22-24, 2000, Maastricht: Universitaire Pers Maastricht, 489-496, 2002c

TAGNIN, S.E.O. Os *Corpora*: instrumentos de auto-ajuda para o Tradutor. In **Cadernos de Tradução**, número especial sobre Corpora e Tradução, Florianópolis, SC: Universidade Federal de Santa Catarina, 191-219. 2002d.

TEIXEIRA, Elisa Duarte. 2005. Compiladora do corpus de Culinária/CorTec: www.fflch.usp.br/dlm/comet.

ULRYCH, Margherita. The impact of multilingua parallel concordancing on translation. In LEWANDOWSKA-TOMASCZYK, B. & MELIA, P.J. (Eds.), **PALC '97 Practical Applications in Language Corpora, Lodz**: Lodz University Press. 421-435, 1997.