UM CORPUS MULTILÍNGÜE PARA ENSINO E TRADUÇÃO – O COMET: DA CONSTRUÇÃO À EXPLORAÇÃO

Stella E. O. Tagnin*

RESUMO: O COMET – Corpus Multilíngüe para Ensino e Tradução –, em construção junto ao CITRAT (Centro Interdepartamental de Tradução e Terminologia) e ao Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, é um corpus eletrônico com vistas a pesquisas lingüísticas principalmente nas áreas de tradução, terminologia e ensino de línguas. O COMET é composto por três sub*corpora*: um Corpus Técnico-Científico, um Corpus de Aprendizes e um Corpus de Traduções. O Corpus Técnico-Científico privilegiará quatro áreas (Direito Comercial, Informática, Ortodontia e Meio Ambiente), e será ampliado sistematicamente até atingir a meta de um milhão de palavras em cada uma das referidas áreas. Abrigará também *corpora* eventuais produzidos pelos alunos do Curso de Especialização em Tradução além de pós-graduandos do programa de Estudos Lingüísticos e Literários em Inglês. Já o Corpus de Aprendizes será constituído de redações em língua estrangeira produzidas por alunos da graduação e dos cursos de extensão das áreas do Departamento de Letras Modernas: alemão, espanhol, francês, inglês e italiano. Finalmente, o Corpus de Traduções subdivide-se em Literário e Juramentado. O Corpus Literário é um corpus paralelo composto de contos traduzidos do inglês, e seus respectivos originais, e de literatura brasileira vertida para idiomas estrangeiros. O Corpus Juramentado será constituído de textos cedidos pela Junta Comercial de São Paulo por meio de contrato de comodato com nossa Universida-

* Universidade de São Paulo.

de. Este artigo relata, justamente, a construção deste último corpus e as possibilidades de pesquisa que oferece.

UNITERMOS: tradução, ensino, pesquisa, fraseologia, terminologia, corpus paralelo, corpus comparável.

ABSTRACT: COMET – A Multilingual Corpus for Teaching and Translation, which is being built at the University of São Paulo, is aimed for linguistic research, especially studies in translation, terminology and language teaching. The COMET consists of three subcorpora: a Technical Corpus, a Learner Corpus and a Translation Corpus. The Technical Corpus favours mainly four areas in which a significant lack of terminological sources has been identified by professional translators: Commercial Law, Computing, Orthodontics and Environmental Sciences. This means that regular work is being carried on to enlarge these corpora systematically to reach 1 million words. In addition, all technical corpora produced by student research or otherwise at the University of São Paulo (USP) will eventually be hosted here. The Learner Corpus will be composed of written assignments by undergraduate and extracurricular students of English, French, German, Italian and Spanish. The Translation Corpus is subdivided into Literary and Sworn translations. The Literary corpus contains American and Canadian short stories and their respective translations into Portuguese, as well as Brazilian literature translated into foreign languages. The Sworn Translations Corpus will consist of sworn translations from and into Portuguese, but no original counterparts. This article reports on COMET's construction and its research possibilities.

KEYWORDS: translation, teaching, research, phraseology, terminology, parallel corpus, comparable corpus

1. Introdução

Atualmente a preocupação lingüística é muito mais descritiva do que prescritiva, ou, nas palavras de Halliday (1966, p.

160), o interesse do pesquisador reside "não apenas naquilo que o falante nativo sabe sobre sua língua, mas também naquilo que com ela faz". Por essa razão, os estudos tendem a se basear em exemplos autênticos, extraídos da linguagem em uso, o que faz dos corpora eletrônicos ferramentas imprescindíveis em diversos âmbitos da pesquisa lingüística, em especial os de natureza lexical.

1.1. Corpora no Brasil

Com o grande o número de corpora informatizados disponíveis, em especial na Europa e, cada vez mais, na Ásia, os estudos empíricos estão reconquistando sua importância na lingüística computacional, dando contribuições significativas para as áreas da aquisição de conhecimento lexical, construção gramatical e tradução automática (Church & Mercer, 1993, p. 2).

No Brasil, no entanto, esses estudos ainda são incipientes. Embora haja diversos corpora em português espalhados pelo país (alguns sediados em universidades, outros de caráter particular), poucos ainda são os trabalhos ensejados por esse material, exceção feita aos estudos baseados no Projeto NURC, cuja forma eletrônica está sediada no Centro de Informática de nossa Faculdade, e aos abaixo discriminados.

Na Universidade Federal do Rio de Janeiro, o **Projeto de Estudos de Usos Lingüísticos (PEUL)**, coordenado por Anthony Julius Naro, também já rendeu um grande número de estudos e publicações.

Na Universidade Estadual Paulista, no *campus* de Araraquara, um corpus de "25.000 páginas de língua escrita" (Borba, 1990) serviu de base para a compilação do **Dicionário Gramatical de Verbos**, coordenado por Francisco da Silva Borba, publicado em 1990.

O mesmo corpus, hoje com 70 milhões de ocorrências, serviu para a compilação do **Dicionário de usos do Português do Brasil** (Borba, 2002) e da **Gramática de Usos do Português** (Neves, 2000). Embora em sua totalidade seja acessível apenas na sede (UNESP/Araraquara), os autores disponibilizam partes dele para determinados projetos de pesquisa.

Outro corpus de língua geral, com aproximadamente 41 milhões de palavras, pertence ao NILC (Núcleo Interinstitucional de Lingüística Computacional) da Universidade de São Paulo, *campus* de São Carlos e é disponibilizado *online* no endereço http://www.linguateca.pt. É formado por textos jornalísticos, didáticos, epistolares, jurídicos, acadêmicos e redações de alunos e vestibulandos. Parte desse corpus (aproximadamente 24 milhões de palavras) compõe o CETENFolha (extratos do jornal Folha de São Paulo do ano 1994), uma contrapartida para o CETEMPublico, um corpus de aproximadamente 180 milhões de palavras em português europeu do jornal diário Público. Ambos estão disponíveis separadamente no site da Linguateca, o que permite pesquisas paralelas. Vale notar que lá se encontram ainda diversos outros corpora, todos, porém, do português europeu.

Outro corpus para o português do Brasil, e também disponibilizado na Linguateca, é o Corpus ECI-EBR, "uma selecção de excertos de obras brasileiras, contendo pelo menos discurso literário, didáctico e oral cuidado (discursos políticos), correspondente a pouco mais de 700 mil palavras" (http://www.linguateca.pt).

Um corpus, da Linguateca, de grande interesse para a área da tradução é o COMPARA. Trata-se de um corpus de excertos literários paralelos – originais e respectivas traduções – nas variantes européia e brasileira do português e britânica, americana e sul-africana do inglês. Possui uma interface bastante amigável, permitindo uma variedade de combinações de parâmetros, que produzem concordâncias paralelas. Não contempla, por ora, textos técnicos, nem permite acesso ao texto como um todo (Frankenberg-Garcia & Santos, 2003).

Na Universidade Federal de Pernambuco está sendo compilado um corpus de variações lingüísticas de Pernambuco, VALPB, sob a coordenação de Luiz Antonio Marcuschi, para a "análise de processos na relação fala e escrita" (Marchuschi, 2001).

Na Pontificia Universidade Católica de São Paulo está sendo construído o Corpus Direct, cujo objetivo "é promover estudos que descrevam eventos discursivos orais e escritos inéditos do âmbito profissional, em português como língua materna e em inglês relevante para o Brasil". Trata-se de "um corpus informa-

tizado de linguagem dos negócios, estruturado e enriquecido com anotação de vários aspectos lingüísticos" (http://lael.pucsp.br/direct/projeto.htm).

Um corpus multilíngüe, multifuncional de discurso, para análise lingüística e literária, denominado CORDIALL, está em construção na Universidade Federal de Minas Gerais (Pagano *et al*, neste volume). É subdividido em três subcorpora e contempla as línguas inglesa, alemã e espanhola, além do português. Dentre os gêneros que engloba, estão principalmente o literário, o acadêmico e o jornalístico.

Além desses *corpora*, há bancos de dados comerciais, disponíveis em CD-ROM, cujas ferramentas de busca, no entanto, são bastante precárias para pesquisas lingüísticas mais elaboradas.

1.2 Corpora: pesquisas

As pesquisas baseadas em corpora têm contribuído sobremaneira para o melhor conhecimento das línguas estudadas, em diversos âmbitos.

No âmbito gramatical, estudos individuais, assim como gramáticas propriamente ditas, demonstraram que muitas das regras das gramáticas normativas não são "obedecidas" na linguagem em uso, seja na sua forma escrita ou oral.

No ensino de línguas, a abordagem DDL (Data Driven Learning – Aprendizado por meio de Dados) (Johns, 1991a e b) privilegia uma abordagem cognitiva do aprendizado, permitindo que o aluno aprenda "por descoberta" e não por memorização. Para tanto, vale-se de concordâncias extraídas de corpora que servem para o aluno fazer inferências a respeito, por exemplo, de determinadas colocações lexicais ou padrões sintáticos em que certo item lexical ocorre.

No âmbito da lexicografia, os corpora não só têm facilitado o trabalho de pesquisa do lexicógrafo, como têm contribuído para que itens antes negligenciados ou mesmo ignorados passassem a fazer parte dos dicionários elaborados segundo os princípios da Lingüística de Corpus (Bowker 1999, 2002; Bowker & Pearson,

2002; Tagnin, 2000, 2002c). Esses itens são, em especial, fraseologias, como as colocações, binômios, frases feitas e outros. Laranjinha (1999), por exemplo, identificou, a partir de um corpus de legislação comercial em inglês e português, colocações verbais peculiares a esse discurso que não constam dos dicionários terminológicos da área.

Esse aspecto é crucial para o tradutor, principalmente quando se trata de áreas técnicas, onde predomina uma terminologia tradicional, que praticamente ignora a fraseologia típica dessas áreas. Para o tradutor é imprescindível não só o contexto de ocorrência de um termo, mas também a fraseologia da área. Dessa forma, os glossários, que vêm sendo desenvolvidos no âmbito do Curso de Especialização em Tradução (CETRAD), incluem tanto os termos quanto a fraseologia de cada área.

No campo da tradução, especificamente, os corpora também têm se mostrado essenciais, a ponto de haver se criado uma área específica denominada Corpus-Based Translation Studies (Estudos de Tradução baseados em Corpora). Nessa área têm se desenvolvido estudos, a partir de textos autênticos, que consideram suas regularidades lingüísticas como normas probabilísticas (Toury, 2000), as quais, por sua vez, refletem variáveis sócioculturais. Outro aspecto que tem sido objeto de estudo são os "universais da tradução" (Baker 1995, 1999; Alves e Magalhães, neste volume), ou seja, procedimentos comuns aos tradutores, independentemente da língua a partir da qual estejam traduzindo, tais como normalização, simplificação e explicitação, que podem ser investigados a partir de certos padrões lingüísticos (Olohan, 2003). Por exemplo, a análise das traduções de *não* no romance *A Hora da Estrela*, de Clarice Lispector, para o inglês – em *The Hour of the Star*, por Giovanni Pontiero – permitiu que Nelia Scott identificasse dois desses padrões de normalização (Laviosa, 2004, neste volume).

Corpora paralelos – originais com suas respectivas traduções – também permitem a investigação de estratégias de tradução, assim como peculiaridades da língua traduzida. Uma metodologia, que denominei "ping-pong" (Tagnin, 2003d), permitiu identificar, com o uso do corpus paralelo COMPARA (http://www.linguateca.pt/COMPARA/), o comportamento de verbos de

elocução em textos originais e traduzidos em português brasileiro e diversas variantes do inglês. O estudo (Grossman *et al.*, 2002) demonstrou que o inglês não se preocupa com a variedade, empregando, na maior parte das vezes, o verbo *say*. Já o português tende a utilizar verbos de elocução mais específicos (*dizer, responder, perguntar, comentar, indagar, exclamar, explicar etc.)*. No entanto, nos textos traduzidos essa característica não ocorre, pois o tradutor, talvez preocupado com a fidelidade ao texto original, acaba por optar por uma tradução literal, produzindo, assim, um texto em "tradutês", como se convencionou denominar.

2. O COMET

Face à enorme possibilidade de pesquisas contrastivas no âmbito do ensino de línguas estrangeiras e da tradução e, principalmente, ao fato de não haver corpora bilíngües de áreas técnicas que envolvam o português brasileiro, decidiu-se pela construção do COMET – um **Co**rpus **M**ultilíngüe para **E**nsino e **T**radução.

O COMET é constituído de três subcorpora: um Corpus Técnico-Científico, um Corpus de Aprendizes e um de Traduções.

2.1 O Corpus Técnico-Científico (CORTEC)

O CORTEC constitui um corpus técnico-científico de âmbito geral, mas, em sua primeira etapa, privilegiará quatro grandes áreas, determinadas a partir de questionário submetidos a diversos tradutores profissionais, via Internet, indagando sobre as áreas mais carentes de material de apoio lingüístico. As respostas apontaram para:

- Direito Comercial
- Informática
- Ortodontia
- Meio Ambiente

A outra parte do corpus já está sendo construída a partir de todos os corpora compilados pelos alunos do Curso e Especialização em Tradução (CETRAD) – Inglês e da Pós-Graduação – resultantes de projetos diversos (Tagnin, 2003a), alguns dos quais serviram para a elaboração de glossários destinados a tradutores (disponíveis em http://www.fflch.usp.br/citrat). Numa primeira etapa (2001), foram construídos corpora com aproximadamente 100.000 palavras em cada língua nas seguintes áreas:

- Biotecnologia: alimentos transgênicos
- Culinária: receitas
- Computação: segurança na Internet
- Moda: roupas
- Veterinária: doenças dos bovinos
- Ecologia: biodiversidade
- Odontologia: ortodontia
- Automação industrial: sensores
- Negócios: mercado financeiro
- Turismo: ecoturismo
- Engenharia genética: genoma

A esses, foram acrescentados, em 2003, 14 outros domínios assim constituídos:

Área	Inglês: arquivos / tokens	Português: arquivos / tokens
Dermatologia	89 - 96.968	62 - 82.781
Esterilização-autoclaves	40 - 73.426	23 - 59.904
Futebol	89 - 96.968	86 - 112.504
Impressoras	72 - 127.482	18 - 10.845 21 Tr. - 102.646
Insuficiência Cardíaca	13 - 10.867	13 - 4.695
Investimento	49 - 182.025	28 - 98.047
Medidores Eletromagnéticos	19 - 61.704	24 - 87.510
Mercado de Capitais	59 - 120.730	44 - 57.470
Nefrologia	45 - 196.451	50 - 185.763
Fundos Oceânicos	30 - 144.411	22 - 85.548
Prostodontia	54 - 184.660	59 - 107.348
Suplementos Nutricionais	133 - 164.004	110 - 129.656
Telecom - Banda Larga	15 - 135.032	21 - 74.896
Desenvolvimento Infantil	16 - 81.834	27 - 129.598
Total	1.680.771 palavras	1.329.206 palavras

Figura 1: Composição dos corpora coletados na 2ª. etapa (2003)

Dessa forma, temos atualmente, em número de palavras,

	Inglês:	Português:
1a. Etapa:	3.782.826	1.599.734
2a. Etapa:	1.680.771	1.329.206
Total:	5.463.597	2.928.940

Em sua grande maioria, os textos constituem corpora comparáveis, ou seja, são de uma mesma área, seguindo padrões semelhantes, tais como: gênero, tipologia textual, extensão, (em geral, ao redor de 100.000 palavras), fonte, data etc., em inglês e português.

2.1.1 Problemas na construção dos corpora

Ao contrário do que se pode imaginar – que basta "baixar" da Internet textos de um determinado assunto para já se obter um corpus – a construção de um corpus, para que atenda os objetivos a que se propõe, deve seguir rigorosos critérios de compilação (Atkins *et al*, 1992). O desconhecimento desses critérios, em especial na primeira etapa do projeto, acarretou uma série de problemas.

2.1.1.1 Delimitação da área de estudo

Um dos primeiro problemas detectados foi o da delimitação da área a ser privilegiada. Assim, embora de início tenha-se pensado, por exemplo, em construir um corpus sobre Informática, logo se percebeu que era imprescindível um enfoque mais restrito, ou seja, abordar uma área bem mais específica, o que levou os pesquisadores a optar, no caso em questão, na primeira etapa, apenas pelo domínio Segurança na Internet.

2.1.1.2 Obtenção dos textos

Pela facilidade de acesso e por já se encontrarem em formato eletrônico, optou-se por privilegiar textos disponíveis na

Internet. No entanto, isso não era possível para todas as áreas: contratos de tipos diversos, que viriam a constituir um subcorpus do Direito Comercial, raramente são encontrados na Internet, de modo que foi necessário obtê-los de outras fontes e, então, escaneá-los ou digitá-los.

A qualidade lingüística dos textos e a idoneidade das fontes também são parâmetros a serem levados em conta para que não haja desperdício de tempo e trabalho, obrigando o pesquisador a descartar material já coletado, o que ocorreu com freqüência maior do que a desejada na primeira etapa.

Além disso, verificou-se certa disparidade entre a quantidade de textos disponíveis nas duas línguas – na maioria das áreas é muito maior o número de textos em inglês. Isso acarretou, em diversos casos, um desequilíbrio entre o volume de textos nas duas línguas (vide Figura 1).

2.1.1.3 Balanceamento do corpus

Para assegurar a comparabilidade dos corpora dentro de uma área, os textos devem: a) inserir-se nos mesmos gêneros e tipos textuais; b) ter extensão semelhante; e c) pertencer ao mesmo período cronológico. No entanto, nem sempre foi possível, dentro do tempo estipulado, obter esse balanceamento entre cada subcorpus das línguas, especialmente porque, em certos casos, a disponibilidade de determinados tipos textuais era maior em uma língua do que em outra. Por exemplo, no caso do corpus de Biodiversidade, há maior número de textos acadêmicos em português do que em inglês. Por essa razão, sugeriu-se, nesse caso, promover uma coleta em paralelo, ou seja, a cada texto ou certo número de textos de determinada tipologia textual compilado numa língua, foi necessário coletar, em seguida, igual número da outra.

2.1.1.4 Padronização dos textos para inserção no corpus

A fim de que os textos possam ser acessados por uma ferramenta de busca e exploração do corpus, em geral é desejado

estarem em formato .txt, o que requer a conversão de formato daqueles que não se encontram assim. No caso de textos obtidos da Internet, após serem baixados no formato .txt, devem ser excluídas todas as marcas de SGML. Essa etapa é de extrema importância para evitar distorção dos resultados quando da investigação do corpus. Assim, se não forem excluídos os elementos extra-textuais, tais como referências bibliográficas e endereços eletrônicos, a lista de palavras existentes naquele texto ou corpus poderá elencar, por exemplo, www com um alto índice de freqüência, em detrimento de outras palavras de maior relevância no corpus em estudo.

2.1.1.5 Inserção do cabeçalho

A última etapa é a inserção do "cabeçalho" para cada texto. Nesse caso, o problema não diz respeito à inserção propriamente dita, mas à concepção do cabeçalho, uma vez que foi preciso um exaustivo estudo sobre gêneros e tipologia textual antes de se chegar a uma configuração adequada. Esse trabalho foi desenvolvido dentro do projeto Lácio-Web, do qual o COMET é parceiro (vide 2.2) e está em constante atualização em função de novos gêneros ou tipos textuais dentre outros parâmetros definidos pelo projeto. Pela sua abrangência, esse cabeçalho se presta a uma grande gama de textos e está disponível para qualquer pesquisador interessado (veja-se www.nilc.icmc.usp.br/lacioweb).

2.1.1.6 Fragmentação das áreas

Esse problema diz respeito ao fato de apenas uma pequena parte das áreas do conhecimento ser privilegiada. Isso decorre do fato de, até o momento, os corpora terem sido compilados segundo as preferências dos pesquisadores, muitas vezes de acordo com seus campos de atuação. O que se pretende, doravante, é mapear cada grande área, determinando-se as subáreas e, a partir daí, buscar construir corpora para esse grupo. É o que já está sendo feito em relação às áreas de Odontologia e Meio Ambiente.

2.1.1.7 Permissão para uso dos textos

Uma vez que se pretende disponibilizar boa parte do COMET na Web, é imprescindível obter-se permissão dos autores e/ou detentores dos direitos autorais para tornar público esse material. Esse é um trabalho árduo e demorado, pois nem sempre é possível identificar a autoria do texto, ou estabelecer contato com o responsável, ou ainda obter a permissão, mesmo em se tratando de material que será utilizado exclusivamente para fins de pesquisa. À medida que se obtenham as permissões, os respectivos textos vão sendo disponibilizados por intermédio do projeto Lácio-Web. Já o corpus de traduções juramentadas tem a vantagem de, por sua própria natureza, constituir-se de textos de domínio público.

2.1.2 A composição atual do CORTEC

O material já coletado foi redistribuído da seguinte forma, dentro das quatro áreas prioritárias:

- **Informática**
  - Segurança na Internet
  - Impressoras
- **Ortodontia**
- **Direito Comercial**
  - Legislação americana/brasileira (Laranjinha, 1999)
- **Meio Ambiente**
  - Ecologia: Biodiversidade
  - Turismo: Ecoturismo

Como há diversos corpora da área de **Medicina**, os seguintes serão agrupados dentro desse domínio:

- Dermatologia
- Insuficiência Cardíaca
- Nefrologia
- Hipertensão Arterial (Castanho, 2003)

Por se tratar de um corpus técnico-científico, é essencial que seja atualizado constantemente, o que implica tanto o acréscimo de textos recentes, para garantir a atualidade da terminologia, quanto a criação de novas áreas ou renomeação de outras, como é o caso do Ecoturismo, hoje denominado Turismo Sustentável.

2.1.3 Um corpus paralelo – Revista Pesquisa FAPESP

Dentro do Corpus Técnico-Científico está também sendo construído um corpus paralelo (originais com respectivas traduções) com os textos eletrônicos da Revista Pesquisa da FAPESP, a partir da edição de número 60 (ano 2000), gentilmente cedidos por aquela instituição. A revista é composta de diversas seções e cobre áreas como Política Científica e Tecnológica, Ciência, Tecnologia e Humanidades. O corpus está sendo alinhado no nível da sentença com o programa TagAlign (Caseli, 2003), em desenvolvimento no Núcleo Interinstitucional de Lingüística Computacional – NILC – da USP/São Carlos. É importante salientar que nesse corpus os textos originais são em português e os traduzidos, em inglês. Além das pesquisas contrastivas lexico-gramaticais de praxe, a diversidade de tipologia textual da revista (reportagens, cartas, notícias, carta do editor, artigos) permitirá estudos também no nível do discurso.

2.1.4 O Projeto Lacio-Web[1]

Os participantes desse projeto são o Núcleo Interinstitucional de Lingüística Computacional (NILC) da USP/São Carlos, o Instituto de Matemática e Estatística (IME/USP) e a Faculdade de Filosofia, Letras e Ciências Humanas/USP, por meio do

[1] Projeto Lacio-Web: Disponibilização de *corpora* e Ferramentas de Navegação e de Busca via Web para Análise Lingüística e Criação de Recursos e Ferramentas Computacionais para o Português – CNPq 09/2001 – SocInfo/ProTeM01/2001, Processo 552176/01-0 (NV) Modalidade AI.

COMET. O projeto prevê a criação de recursos computacionais e lingüísticos de base (corpora e suas ferramentas associadas) para a implementação de ferramentas de processamento automático do português, necessárias para a criação, organização, manipulação e busca de informação em português na Web. Para tanto, já está disponibilizando diversos corpora via Web com ferramentas de acesso para análises lingüísticas (www.nilc.icmc.usp.br/lacioweb).

O COMET tem uma participação significativa, contribuindo com uma parte de seu Corpus Técnico-Científico em português para a compilação do corpus principal do Lácio-Web, o Lácio-Ref, um corpus composto de textos autênticos, de diversos gêneros e tipos textuais, e que pretende ser um corpus de referência do português brasileiro, acessível via Web. Contribui ainda com textos paralelos para o desenvolvimento de alinhadores. Em contrapartida, tem acesso às ferramentas que estão sendo desenvolvidas dentro do projeto, visando, principalmente, à possibilidade de estudos lexicais de variada espécie, estudos tradutológicos, bem como a construção de glossários de especialidade.

2.2 O Corpus de Aprendizes

Esse corpus, por não estar diretamente relacionado à tradução – tema deste número especial –, não será aqui descrito. Para maiores detalhes, vide Tagnin, 2003c.

2.3 O Corpus de Traduções

Esse corpus é constituído de dois subcorpora: a) Traduções Literárias, e b) Traduções Juramentadas.

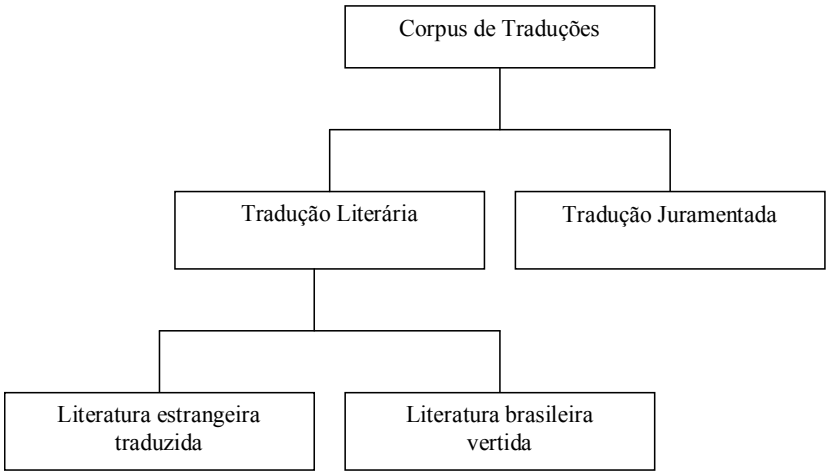

Figura 2: Constituição do Corpus de Traduções

O subcorpus de Traduções Literárias compreende dois conjuntos com materiais já coletados, resultantes de outros projetos em andamento:

a. um conjunto de textos paralelos (originais e respectivas traduções) de literatura estrangeira traduzida para o português brasileiro, que consiste de nove contos americanos e vinte contos canadenses traduzidos por alunos do CETRAD. Os contos canadenses foram publicados em 2002, sob o título ***Lá do Canadá*** (Tagnin, 2002d). Em 2004 o corpus será acrescido de aproximadamente 25 contos australianos e suas respectivas traduções;
b. um conjunto de textos paralelos de literatura brasileira vertida para idiomas estrangeiros.

O corpus de Traduções Juramentadas será composto de textos extraídos dos livros de registro dos tradutores juramentados de São Paulo, disponibilizados para o Centro Interdepartamental de Tradução e Terminologia (CITRAT)/USP pela Junta Comercial do Estado de São Paulo (JUCESP) por meio de um Contrato de Comodato. O material cobre um período de 100 anos (1902-2002), abrangendo cerca de 20 línguas. Para viabilizar e agilizar a cons-

trução desse corpus, decidiu-se digitalizar, por ora, apenas os textos nas cinco línguas do Departamento – alemão, espanhol, francês, inglês e italiano, a partir da década de 60 (Aubert & Tagnin, 2003, e Aubert & Tagnin, neste volume).

2.4 Caracterização do COMET

2.4.1 Os objetivos

Face ao acima exposto, tornam-se claros os objetivos desse corpus no campo da tradução. Em primeiro lugar, o COMET pretende ser uma fonte de linguagem natural atualizada em diversas áreas técnico-científicas. Com a configuração descrita, pretende compensar a falta de material lexicográfico e terminológico nas áreas contempladas. Acima de tudo, no entanto, pretende ser fonte representativa para a pesquisa, a prática e o ensino da tradução, como já vem ocorrendo.

2.4.2 O público-alvo

O público a que se destina abrange desde aprendizes e professores de tradução, tradutores nas duas direções (inglês e português), até lexicógrafos e terminólogos, além de quaisquer pesquisadores interessados nos vários aspectos lingüísticos desses idiomas, inclusive na análise do discurso.

2.4.3 Os textos

Em virtude dessa abrangência, os textos que compõem o COMET inserem-se, na sua grande maioria, em três gêneros: acadêmico, jornalístico e comercial.

Os textos acadêmicos são aqueles escritos por especialistas para especialistas. Caracterizam-se por apresentarem a linguagem natural empregada por esses profissionais, ou seja, apresentam o termo em seu contexto natural, inclusive com suas colocações e coligações. Esse aspecto é essencial para o tradutor

que, com freqüência, tem dúvidas quanto às palavras (verbos, adjetivos) que co-ocorrem com o termo em questão.

Os textos jornalísticos nas áreas técnico-científicas são, em geral, escritos por especialistas para um público leigo. Por essa razão, apresentam muitas vezes uma definição dos termos técnicos, aspecto de especial interesse para o terminólogo. O tradutor, porém, também se beneficia desse tipo de texto, pois o contexto de ocorrência pode assegurar-lhe a equivalência (ou não) de um termo sobre o qual esteja em dúvida.

Finalmente, os textos comerciais (folhetos, manuais, anúncios etc.), escritos por especialistas ou não-especialistas para um público leigo, são de grande valia pela alta concentração de termos técnicos e, muitas vezes, pelas ilustrações que os acompanham, o que contribui para esclarecer o significado de termos obscuros.

É essa, enfim, a configuração que norteia a coleta dos textos das áreas que compõem o COMET.

É preciso também mencionar que os textos são inseridos na íntegra, não só para assegurar a possibilidade de análise textual, como também para servirem de fonte de referência para o estudo do assunto tratado. O público que mais se beneficia desse aspecto são os aprendizes de tradução, que podem, dessa forma, familiarizar-se com o assunto em que estão trabalhando. É fato que, quanto maior o conhecimento de uma área, mais apto está o aprendiz para produzir uma tradução confiável.

2.4.4 As línguas

O CORTEC está sendo coletado apenas nos idiomas inglês e português. Já o Corpus de Traduções abarca também outras línguas, em especial francês, alemão, espanhol e italiano.

2.4.5 Tipos de corpora

- *Corpora comparáveis*

Conforme mencionado, o CORTEC é um corpus comparável, ou seja, é composto de textos originais nas duas línguas.

Essa composição permite observar o uso natural da linguagem, fornecendo ao tradutor subsídios para produzir uma tradução fluente e natural. Permite também avaliar a equivalência de significado e de uso de um termo ou palavra pela análise do seu contexto de ocorrência e, assim, produzir traduções naturais e glossários bilíngües confiáveis. Ao pesquisador, permite estudos sobre a fraseologia da área (recorremos aqui à Culinária, a título de exemplo), aspecto praticamente ignorado na grande maioria dos glossários. Esses, em geral, se atêm a termos simples (*pimenta, caçarola)* e compostos (*pimenta-do-reino, pimenta calabresa, panela de pressão*), não registrando unidades de significado mais extensas como, por exemplo, *pimenta-do-reino moída na hora*, ou colocações verbais como *untar uma forma, cortar (uma cebola) em rodelas.*

- *Corpora paralelos*

O COMET contém dois corpora paralelos: o da Revista Pesquisa da FAPESP e o Literário. Por corpora paralelos entendemos textos originais e suas respectivas traduções, em uma ou mais línguas.

O corpus FAPESP contém textos originalmente escritos em português e suas traduções para o inglês. Já o corpus Literário subdivide-se em Literatura Estrangeira Traduzida e Literatura Brasileira Traduzida.

Esse tipo de corpus permite analisar processos e estratégias de tradução, bem como enseja toda sorte de estudos contrastivos, desde morfológicos, sintáticos e lexicais até textuais (Schmied; Johansson; Aijmer, neste volume).

- *Corpus de traduções*

Ao contrário dos corpora anteriores, que envolvem apenas textos originais ou originais e respectivas traduções, o Corpus de Traduções Juramentadas compõe-se exclusivamente de traduções, à semelhança do Translational English Corpus (TEC), compilado no UMIST, Universidade de Manchester (Laviosa, 2004, neste volume). Difere dele, porém, por incluir traduções de diver-

sas línguas estrangeiras para o português, bem como do português para diversas línguas estrangeiras, ou mesmo de uma língua estrangeira para outra, ao passo que o TEC se restringe a traduções para o inglês. Sua configuração peculiar – diversidade tipológica, diversidade de línguas, caráter diacrônico (cobre um período de 100 anos) (Aubert & Tagnin, neste volume) – presta-se a todo tipo de estudos lingüísticos e terminológicos, bem como históricos e sócio-políticos, pois a "estreita vinculação entre a tradução pública e as esferas política, institucional, econômica e jurídica sugere que um material que abarca todo um século deverá também portar as marcas dos processos históricos vivenciados pela comunidade em que tais traduções foram executadas e recebidas" (Ibidem p. 167).

3. Pesquisas em andamento

O COMET, em seu estágio atual, já está sendo usado para pesquisas em várias áreas:

3.1 Construção de Corpora

Com base nos critérios estabelecidos por Atkins et al (1992), está sendo analisado o corpus de Ortodontia no intuito de identificar problemas na sua construção inicial e sugerir novos parâmetros para garantir que sua ampliação resulte num corpus confiável e minimamente representativo. O objetivo da pesquisa é que esses novos parâmetros sejam aplicados a todos os corpora que fazem parte do COMET (Pardo 2004).

Uma das áreas prioritárias do COMET – a de Meio Ambiente – está sendo construída a partir dos corpora de Biodiversidade e de Ecoturismo já compilados. Para assegurar que todos os ramos que compõem essa área estejam representados, faz-se necessário estabelecer uma árvore de domínio, cuja definição está em andamento. Uma vez estabelecido o corpus, será extraída a terminologia referente à área do Turismo Sustentável.

Como desdobramento do corpus de Culinária, foi construído um corpus multivarietal de receitas nas variantes português brasileiro e europeu, inglês britânico e americano, cujo objetivo era investigar o quão expressivas são as diferenças entre essas variantes (Tagnin & Teixeira, no prelo). Os resultados apontaram, além de diferenças lingüísticas, também algumas diferenças culturais, o que sugeriu a ampliação do corpus para permitir a identificação dessas características. Estão sendo investigados os critérios para se obter um corpus comparável dessas variantes no nível "cultural".

3.2 Terminologia/Fraseologia

A terminologia tradicional tem privilegiado termos simples e compostos, pouca atenção dando a unidades multipalavras, ou seja, unidades de significado maiores, dentre as quais contam-se os diversos tipos de colocações, os binômios, as coligações, as expressões idiomáticas etc. Essas compõem a dimensão convencional da língua, ou seja, sua fraseologia. O corpus de Direito Comercial, em inglês e em português, está sendo investigado para se detectar uma dessas unidades – os binômios ,– categoria de presença marcante no discurso jurídico.

As colocações são objeto de outra pesquisa, dessa feita no âmbito do "Business English". Essa pesquisa tem o intuito de identificar as colocações mais recorrentes num corpus de textos autênticos da área para estabelecer o vocabulário a ser ensinado a aprendizes do inglês como língua instrumental na área dos negócios. O estudo vem demonstrando que o vocabulário constante nos livros-texto da área não reflete, no geral, o vocabulário de fato empregado pelos profissionais (Orenha, 2004).

3.3 Lexicografia

Está em desenvolvimento um projeto para "Vocabulários eletrônicos bilíngües: proposta de consulta modular para tradutores". Como o tradutor-alvo é o tradutor técnico, essa pesquisa

está se baseando em cinco subcorpora do COMET, a saber: Informática, Culinária, Ortodontia, Direito Comercial e Ecoturismo.

A criação de um Dicionário Fraseológico Bilíngüe de Culinária é outro projeto em andamento. Constituirá fonte de referência específica para tradutores e, como tal, suas entradas privilegiarão as informações necessárias a esse profissional, em especial, o contexto de ocorrência do termo ou item fraseológico e seu equivalente "funcional" na língua de chegada.

4. Conclusão

O COMET é um corpus multilíngüe destinado ao ensino e à pesquisa de línguas e de tradução. Neste artigo enfocamos em especial sua relevância para a tradução e a terminologia. Sua subdivisão em três corpora distintos – um Corpus Técnico-Científico (CORTEC), composto de textos originais em inglês e em português, um Corpus de Aprendizes (não discutido aqui) e um Corpus de Traduções, composto de traduções literárias e traduções juramentadas – facilita sua utilização para a resolução de questões práticas, tais como determinar o uso correto de certo termo, ou a palavra que usualmente co-ocorre com outra. Presta-se, outrossim, para uma gama extremamente variada de estudos acadêmicos. Atualmente, no âmbito de nossa Universidade, está sendo usado para trabalhos sobre lexicologia, terminologia, construção de corpora, processos e estratégias de tradução e análises contrastivas. Outros trabalhos envolvem a construção de corpora próprios os quais, ao término, serão incorporados ao COMET. Dessa forma, estabelece-se uma proveitosa troca acadêmica: o COMET alimenta diversos estudos acadêmicos ao mesmo tempo em que corpora resultantes de outros estudos retroalimentam o COMET. Assim, garante-se o enriquecimento e a constante atualização do corpus.

Referências bibliográficas

ALUÍSIO, S. M., PINHEIRO, G.M., FINGER, M. NUNES, M.G.V. e TAGNIN, S.O. (2003) The Brazilian Portuguese Corpus Lacio-Web: Overview and issues in corpus creation**.** In Archer, Dawn, Paul Rayson, Andrew Wilson and Tony McEnery (eds.), *Proceedings of the Corpus Linguistics 2003*. Lancaster University (UK), UCREL Technical Papers, v. 16, part 1, Special Issue, p. 14-21; ISBN 1 86220 131 5.

ATKINS, S, CLEAR, J. & OSTLER, N. (1992) Corpus Design Criteria, *Literary and Linguistic Computing*, vol. 7, n. 1, p. 1-16.

AUBERT, F.H. & TAGNIN, S.E.O. (2003) A corpus of sworn translations – for linguistic and historical research. In Archer, Dawn, Paul Rayson, Andrew Wilson and Tony McEnery (eds.), *Proceedings of the Corpus Linguistics 2003*. Lancaster University (UK), UCREL Technical Papers, vol. 16, part 1, Special Issue, p. 54-61; ISBN 1 86220 131 5.

AUBERT, F.H. & TAGNIN, S.E.O. (2004) Um corpus de traduções juramentadas – material de pesquisa lingüística, sociológica e histórica, neste volume.

BAKER, M (1999). The Role of Corpora in Investigating the Linguistic Behavior of Professional Translators. *International Journal of Corpus Linguistics*, vol. 4(2), p. 281-98.

_____. (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target* **7**, John Benjamins, p. 223-43.

BORBA, F. da Silva (coord.) (1990) *Dicionário Gramatical de Verbos do Português Contemporâneo*. São Paulo: UNESP.

_____. (2002). *Dicionário de Usos do Português do Brasil*. São Paulo: Ática.

BOWKER, L. (1999). Exploring the Potential of Corpora for Raising Language Awareness in Student Translators. *Language Awareness,* v. 8, n. 3&4, p. 160-73.

_____. (2002) *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

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

CASTANHO, R.M.C. (2003) *Proposta para a elaboração de um glossário de colocações na área médica – subárea hipertensão arterial*. Dissertação de mestrado, Universidade de São Paulo.

CHURCH, K. W. & MERCER, R.L. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora, *Computational Linguistics* 19: 1, p. 1-24.

FRANKENBERG-GARCIA, A. & SANTOS, D. (2003). COMPARA, um corpus paralelo de português e inglês na Web. *Cadernos de Tradução* no. IX – 2002/1, número especial sobre Corpora e Tradução. Universidade Federal de Santa Catarina, Florianópolis: Núcleo de Tradução, p. 61-79.

HALLIDAY, M.A.K. (1966). Lexis as a Linguistic Level. In BAZELL, D.E., CATFORD, J.C., HALLIDAY, M.A.K. & ROBINS, R.H. (org.). *In Memory of J. R. Firth*. London, p. 148-62.

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

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

LARANJINHA, A.L.T. (1999) *Para um Glossário Bilíngüe – Português/Inglês de Termos do Direito Comercial: Colocações Verbais.* Dissertação de mestrado, Universidade de São Paulo.

LAVIOSA, S. (2004) Corpus-based Translation Studies: Where does it come from? Where is it going?, neste volume.

MARCUSCHI, L.A. (2001) Um corpus lingüístico para a análise de processos na relação fala e escrita. Ms. apresentado no XI InPla, PUC-SP, 4-5/05/2001.

NEVES, M.H.M. (2000) *Gramática de Usos do Português.* São Paulo: Editora Unesp.

OLOHAN, M. (2003) Leave it out! Using a comparable corpus to investigate aspects of explicitation in Translation. In *Cadernos de Tradução* no. IX – 2002/1, número especial sobre Corpora e Tradução. Universidade Federal de Santa Catarina, Florianópolis: Núcleo de Tradução, p. 153-69.

ORENHA, A. (2004) O inglês dos negócios: quais termos ensinar? (dissertação de mestrado, título provisório).

PAGANO, A, MAGALHÃES, C.M. & Alves, F. (2004) Towards the construction of a multilingual, multifunctional corpus: factors in the design and applications of CORDIALL, neste volume.

PARDO, R. M. (2004) Critérios de Construção e Organização de um Corpus de Especialidade: o Corpus Técnico-Científico de Ortodontia. Dissertação de mestrado, FFLCH/USP.

SINCLAIR, J. M. (2002) Corpus-driven Linguistics. 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, Gotemburgo.

TAGNIN, S.E.O. (2000) Collecting data for a bilingual dictionary of verbal collocations: From scraps of paper to corpora research. In LEWANDOWSKA-TOMASZCZYK, B. e MELIA, P.J. (eds.) *PALC '99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Lodz.* Frankfurt am Main: Peter Langh GmbH, p. 399-407.

_____. (2001) COMET – A Multilingual Corpus for Teaching and Translation, Comunicação apresentada na PALC '01 – International Conference on Practical Applications in Language Corpora. Lodz.

_____. (2002a) Um corpus de integração: o projeto COMET. Comunicação apresentada no 5. Simpósio de Lingüística de Corpus, 2002, Criação, anotação e aplicação de corpora (InPLA, PUC/SP). São Paulo.

_____. (2002b) Taking off in Brazil: COMET – A Multilingual Corpus for Teaching and Translation. Comunicação apresentada no ICAME 2002 – The Theory and Use of Corpora – The 23rd International Conference on English Language Research on Computerized Corpora of Modern and Medieval English, Gotemburgo.

_____. (2002c) Corpora and the Innocent Translator: How can they help him, in Thelen, Marcel (ed.) *Translation and Meaning, Part 6***,** Proceedings of the Lodz Session of the 3rd Maastricht-Lodz Duo Colloquium on "Translation on Meaning". Lodz, Maastricht: Universitaire Pers Maastricht, p. 489-96.

_____. (org.) (2002d) Lá do Canadá – contos. São Paulo: Olavobrás.

_____. (2003a) Os Corpora: instrumentos de auto-ajuda para o Tradutor. In *Cadernos de Tradução* no. IX – 2002/1, número especial sobre Corpora e Tradução. Universidade Federal de Santa Catarina, Florianópolis: Núcleo de Tradução, p. 191-219.

_____. (2003b) "Ping-Pong" – uma metodologia para análise contrastiva baseada num corpus paralelo, no simpósio *Lingüística de Corpus: Descrição e Tradução* do 13º. InPLA (Intercâmbio de Pesquisas em Lingüística Aplicada), São Paulo (PUC-SP).

_____. (2003c) A multilingual learner corpus in Brazil, In: ARCHER, D., RAYSON, P., WILSON, A. e MCENERY, T. (eds.), *Proceedings of the Corpus Linguistics 2003.* Lancaster University (UK), UCREL Technical Papers, vol. 16, part 1, Special Issue, p. 940-5; ISBN 1 86220 131 5, A ser publicado também em WILSON, A. RAYSON, P. e ARCHER, D. (eds.) (no prelo) Corpus Linguistics around the world (da série Language and Computers). Rodopi, Amsterdam.

TOURY, G. (2002) The Nature and Role of Norms in Translation. In: VENUTI, L. (ed.) *The Translation Studies Reader*. London/New York: Routledge (artigo originalmente escrito em 1978 e revisto em 1995).