Idéias que Cruzam o Oceano[1]

Belinda Maia e Luís Sarmento[*]
Stella E. O. Tagnin e Sandra Maria Aluísio[**]

Resumo: *O objetivo deste artigo é apresentar projetos similares e, por vezes, complementares implementados ou por implementar nos dois lados do Oceano Atlântico: na Universidade do Porto, em Portugal e na Universidade de São Paulo, no Brasil. Para tanto, subdivide-se em duas partes: na primeira, os autores portugueses (Maia e Sarmento) detalham e discutem ferramentas desenvolvidas para um publico de tradutores e terminólogos. O Corpógrafo é um conjunto de ferramentas para a contrução e análise de corpora e bases de dados terminológicos, enquanto o TrAva tem por objetivo avaliar tradutores automáticos. No início da segunda parte, Aluísio descreve o portal Lácio-Web, sua construção, os córpus de que se compõe, bem como as ferramentas disponíveis para sua exploração e montagem de sub-córpus de pesquisa. Na parte final, Tagnin descreve o projeto COMET, sua construção, seus vários subcórpus (Cortec, Córpus de Traduções e Córpus de Aprendizes) e os tipos de pesquisa a que se destinam.*

Palavras-chave: *córpus; tradução; terminologia; ferramentas eletrônicas; Corpógrafo; TrAva; Lácio-Web; COMET.*

---

[1] Este trabalho foi originalmente apresentado no Workshop Luso-Brasileiro no Congresso da European Society for Translation Studies (EST), realizado em Lisboa, de 26 a 29 de setembro de 2004.

[*] Universidade do Porto.
[**] Universidade de São Paulo – USP.

Abstract: *This article presents similar and sometimes complementary projects developed and under development across the Atlantic: at the Porto University in Portugal and at the University of São Paulo, in Brazil. It is divided into two parts: in the first one the Portuguese authors (Maia and Sarmento) discuss the computational tools that have been developed for translators and terminologists: the Corpógrafo is a set of tools for building and analyzing corpora and terminological databases, while TrAva aims at evaluating automatic translation software. The second part starts with Aluísio's description of the Lácio-Web portal: its construction, the corpora that constitute it as well as the tools available for investigating and creating personal sub-corpora for research. In the last part Tagnin describes the COMET Project, its construction, its various subcorpora and the types of research these are meant for.*

Keywords: *corpus; translation; terminology; computational tools; Corpógrafo; TrAva, Lácio-Web; COMET.*

Parte 1 – Universidade do Porto, Portugal

0 Introdução

Num congresso que favorece o estudo de tradução do ponto de vista académico/literário/cultural, é possível que a proposta de falar em pesquisa académica apoiada em ferramentas não seja muito popular. No entanto, o estudo das línguas em geral, e da tradução em particular, têm muito a ganhar com o tipo de observação possibilitado pela digitalização de qualquer tipo de texto e pelo respectivo exame cuidadoso com as ferramentas produzidas pela lingüística computacional e pelo processamento da linguagem natural.

Para quem desconhece essa área de pesquisa, lembramos que é costume chamar a um conjunto de textos digitalizados um 'corpus', sendo o plural 'corpora'. Em certos casos, é possível enriquecer um determinado *corpus* com informação adicional recorrendo a etiquetadores que inserem informação no texto. Essa informação é normalmente de natureza lingüística, sendo também possível incluir etiquetas morfossintácticas, utilizando ferramentas como o etiquetador "Palavras" de Eckhard Bick – ver http://visl.hum.sdu.dk/visl/pt/ – utilizado pelo projecto Linguateca, ao qual pertence o PóloCLUP, da Universidade do Porto. O tipo de etiquetagem efectuado depende do objectivo do projecto sendo possível a criação de um novo esquema de etiquetagem, conforme a utilização prevista para o *corpus*.

No entanto, enquanto que os etiquetadores morfossintácticos já funcionam no nível automático ou semi-automático, qualquer etiquetagem de nível semântico ou textual terá ainda de ser feita manualmente.

Uma vez construído o *corpus*, etiquetado ou em 'cru', há ferramentas para observar e analisar o uso da língua nesse *corpus*. Essas ferramentas podem oferecer informação de natureza estatística, que poderá posteriormente ser analisada para objectivos específicos. A ferramenta mais conhecida, entretanto, é a que permite pesquisar concordâncias, i.e., a observação de fenómenos lingüísticos no nível da frase ou do contexto. Essa pesquisa pode ser feita com uma ou várias palavras, ou, no caso de um *corpus* etiquetado, utilizando as etiquetas, e permite ao pesquisador confirmar os seus palpites com exemplos concretos e quantificáveis.

Os grandes corpora monolingües oferecem a possibilidade de estudar a língua no nível lexical e sintáctico. O British National Corpus (BNC), que é composto por uma grande variedade de textos, escritos ou transcritos de textos orais, contendo 100 milhões de palavras, todas etiquetadas no nível de 'parts-of-speech', está disponível para uso simples (até 50 concordâncias) em linha via http://sara.natcorp.ox.ac.uk/lookup.html ao público em geral. A Linguateca – http://www.linguateca.pt – dispõe de corpora monolingües de mais de 250 milhões de palavras com quase 4 milhões anotados, livremente acessível em linha.

Há corpora bi- e multilingües que servem para o estudo comparativo das línguas em todos os níveis, desde o léxico ao texto. Esses corpora podem ser paralelos – os originais e as suas traduções – alinhados normalmente no nível da frase, ou comparáveis – textos originais em duas ou mais línguas, considerados representativos de determinados géneros, registos ou domínios. A Linguateca oferece um *corpus* paralelo em português e inglês de textos literários com aproximadamente 1 milhão de palavras em cada língua.

1. Corpora e Tradução

Actualmente o uso de corpora electrónicos para a pesquisa e ensino de tradução vai muito além de uma simples consulta lexicográfica. As ferramentas disponíveis, aliadas a uma preparação mais sofisticada do utilizador, permitem a análise sintáctica do texto e do discurso. É evidente que a informática obriga o tradutor a observar uma certa linearidade na tradução, o que vai contra as teorias que advogam uma interpretação do texto original adequado ao contexto do texto de chegada. Isto, em si, é um assunto para investigação. Mas também é pertinente considerar se essas teorias mais 'abrangentes' de facto se aplicam mais ao texto literário ou ao texto não literário. A criação de corpora comparáveis permite a observação de textos originais em duas ou mais línguas, dando uma

maior possibilidade de estudar as diferenças entre as convenções textuais em todos os níveis lingüísticos e culturais, e torna-se evidente que as diferenças existem em todo o tipo de texto (Maia, 2003a).

O uso comercial de corpora paralelos, como memórias de tradução, criou um novo paradigma nos estudos de tradução. No nível comercial é preciso melhorar, uniformizar e re-utilizar textos de uso diário. É costume falar de manuais de instruções nesse contexto, porque é óbvio que esses textos precisam de ser bem formulados, seguindo regras estudadas e formuladas pela disciplina de 'Escrita Técnica', ou 'Technical Writing', que as traduções têm de obedecer a níveis de qualidade comparáveis, e que é de interesse económico aproveitar tanto os textos como as traduções já feitas quando se tratam de ligeiras modificações ou aumentos dos textos para máquinas mais desenvolvidas. Entretanto, as memórias de tradução têm grande utilidade para a cuidadosa preparação de textos e traduções em qualquer área que imponha um alto nível de exigência estilística e terminológica. Também têm implicações para a tradução automática

Enquanto a lingüística computacional e a lingüística de corpora evoluíam bastante durante os anos 90, profissionais ligados a essas áreas, mas com responsibilidades na educação de tradutores, começaram a compreender o alcance da aplicação dessas tecnologias ao ensino da lingüística contrastiva e da tradução em geral. Com o advento da Internet e o acesso a uma grande quantidade de informação em todos os domínios, também surgiram novas possibilidades para melhorar o ensino da tradução especializada. As actas dos congressos de TALC – Teaching Applications and Language Corpora (1994, 1996, 1998, 2000, 2002 & 2004), PALC – Practical Applications of Language Corpora (1997, 1999, 2001, & 2003), e CULT – Corpus Use and Learning to Translate (1997, 2000, & 2004) demonstram vários aspectos dessas tendências.

É natural que o mundo da língua portuguesa tenha sido influenciado por essas correntes. Este trabalho vai apresentar e discutir projectos complementares em andamento nos dois lados do oceano, ou seja, nas Universidades do Porto (Portugal) e de São Paulo (Brasil). São projectos que se inserem nos Estudos de Corpora e têm por objectivo desenvolver corpora para diversos fins, assim como ferramentas de exploração desse material.

2. Corpora e tradução – ensino e pesquisa no PoloCLUP e LINGUATECA

Há vários anos, o projecto Linguateca se preocupa com a função de corpora no ensino e pesquisa de lingüística e tradução. Além de uma vasta selecção de textos em português e um forte leque de ferramentas de pesquisa lingüística, o seu *corpus* paralelo COMPARA está disponível em linha. Esse *corpus* foi iniciado em 1999 e, devido essencialmente aos esforços das duas autoras, continua a ser o maior

corpus paralelo revisto, com textos em português e inglês e as suas traduções, e com um elevado número de utilizadores em escala mundial (Frankenberg-Garcia & Santos, 2002 & 2003). É disponibilizado através do DISPARA (Santos, 2002), que oferece capacidades inovadoras de procura e tem uma interface rigorosamente paralela nas duas línguas.

Na Universidade do Porto, o interesse em corpora, lingüística contrastiva e terminologia já resultou em teses de doutoramento e mestrado e artigos de vária ordem desde 1994. Antes de pertencer ao projecto Linguateca, já tínhamos experimentado trabalhar com corpora paralelos e comparáveis e, especialmente, com *'do-it-yourself corpora'* (Maia, 1997) ou *'disposable corpora'* (Varantola, 2002). Quando a Internet começou a oferecer a possibilidade de adquirir informação sobre uma grande variedade de domínios especializados e, ao mesmo tempo, facilitou a utilização de textos em formato digital, os já interessados em corpora não tardaram em compreender as implicações pedagógicas dessas fontes de informação. Rapidamente desenvolveu-se uma metodologia para estudar os textos normalmente associados com essas áreas para o estudo de todos os níveis linguísticos, desde o lexema ao discurso. O nosso interesse em domínios especializados levou a um estudo mais aprofundado da terminologia (Maia, 2003: vários) e à criação de um Mestrado em Terminologia e Tradução na Faculdade de Letras da Universidade do Porto (FLUP), que funciona desde 2000.

3. O Corpógrafo

Em outubro de 2002, o PoloCLUP da Linguateca foi inaugurado e, entre outras coisas, tem desenvolvido pesquisa no uso de corpora especializados comparáveis para o estudo e a extração de terminologia. Criámos, para esse efeito, o Corpógrafo, um conjunto de ferramentas disponível 'online' para quem estiver interessado em pesquisar autonomamente. O Corpógrafo permite coleccionar textos em vários formatos, formar e analisar corpora, extrair terminologia e criar bases de dados terminológicas com vários campos, inclusive com a possibilidade de criar relações semânticas e ontologias.

O Corpógrafo (Sarmento & Maia, 2003; Sarmento, Maia & Santos, 2004) é uma plataforma de pesquisa sobre corpora que surgiu da necessidade de integrar no mesmo ambiente todo um conjunto de operações e de processos que normalmente são realizados utilizando várias ferramentas ou sistemas, cujo acesso é muitas vezes restrito ou difícil. O Corpógrafo oferece ao utilizador, através de uma simples interface Web, a possibilidade de compilar e pesquisar os seus próprios corpora (a partir de documentos em formato PDF, Ms-Word, PostScript, RTF ou HTML) sem que para isso seja necessário ter conhecimentos especiais

de informática. De certa forma, o Corpógrafo complementa a oferta de corpora genéricos já oferecidos pela Linguateca (projectos AC/DC, CETEMPúblico e CETENFolha), possibilitando a construção e pesquisa em corpora pessoais e específicos, por utilizadores com interesses na área da Lingüística, Tradução ou Engenharia do Conhecimento.

Do ponto de vista mais específico dos estudos na área da Lingüística, o Corpógrafo possibilita pesquisas de concordância e colocações, e a realização de estudos de frequências de combinações de duas ou mais palavras sobre os corpora pessoais. Para tarefas associadas à Tradução e à Engenharia do Conhecimento, o Corpógrafo possui também funcionalidades de pesquisa terminológica. A pesquisa terminológica encontra-se directamente integrada num sistema de base de dados para uma fácil organização dos termos extraídos. As capacidades de pesquisa terminológica (fundamentalmente em português e inglês, mas também em espanhol, francês, italiano e alemão) são complementadas com funcionalidades que permitem a identificação de definições dos termos extraídos e de possíveis relações semânticas (actualmente meronímia e hiponímia) entre os conceitos. Actualmente o Corpógrafo foi experimentado por cerca de 200 pessoas e é utilizado regularmente por 40, localizadas maioritariamente em Portugal e no Brasil, sendo também utilizado por investigadores dispersos por vários países da Europa.

Na FLUP, há várias teses de doutoramento e mestrado e projectos de terminologia em curso utilizando o Corpógrafo para pesquisar áreas de Engenharia Mecânica, Engenharia Electrónica, Geografia da População, Geografia – Riscos Naturais, Genética, Neuroanatomia, GPS – Geographical Positioning System e outros. Podemos prever que, com a entrada numa nova fase de funcionamento do Corpógrafo e a continuação do seu desenvolvimento com a pesquisa em curso, poderemos dar um salto qualitativo e quantitativo no nível de trabalho a ser efectuado.

4. Avaliação de Tradução Automática – ensino e pesquisa

Uma outra experiência no âmbito do PoloCLUP e do Mestrado em Terminologia e Tradução surgiu com a avaliação de tradução automática (TA). Essa experiência começou mais no nível pedagógico, embora a tecnologia criada para esse efeito tenha vindo a ser a aproveitada por uma grande variedade de utilizadores.

A primeira ferramenta criada foi o METRA, que permite pedir traduções Português > Inglês e Inglês > Português a sete motores de tradução automática livremente disponíveis na web. A segunda ferramenta foi o BOOMERANG, que permite que cada tradução automática seja re-introduzida e re-traduzida automa-

ticamente até chegar a um ponto de paragem. Embora o resultado tenha um aspecto quase cómico, a idéia subjacente é teoricamente interessante. A terceira ferramenta chamou-se EVAL e resultou no TrAva, uma ferramenta que permite a observação e análise de vários motores de TA online. Um dos resultados é um *corpus* de traduções + traduções automáticas EN > PT, o CorTA. Essas ferramentas estão disponíveis em http://www.linguateca.pt.

O TrAva é uma ferramenta construída essencialmente para a criação de material de teste e foi desenvolvida como proposta inicial de avaliação conjunta para a tradução automática (TA). O TrAva permite traduzir frases do inglês para o português em quatro motores de TA disponíveis na Internet e apresenta um quadro de classificação de critérios lingüísticos utilizando dois sistemas gramaticais: o sistema de etiquetação gramatical utilizado pelo British National Corpus (BNC) para a classificação das frases da língua de partida (inglês) e uma taxonomia baseada na sintaxe do português para a classificação dos resultados na língua de chegada. Os resultados da avaliação de traduções desenvolvidos com o TrAva são consultáveis através do *corpus* CorTA, Corpus de Traduções automáticas Avaliadas (Santos et al., 2004).

O TrAva é um serviço web que tem como objectivo permitir a recolha cooperativa de exemplos de tradução automática (i.e. frases originais e as respectivas traduções automáticas) de forma a criar uma colecção de casos que permita uma melhor compreensão dos problemas envolvidos na TA da língua portuguesa. O modo de funcionamento do TrAva é simples e próximo do serviço ao qual sucede, o EVAL (http://poloclup.linguateca.pt/ferramentas/eval/). O TrAva envia uma frase em inglês fornecida pelo utilizador a quatro serviços de tradução automática web para que esses a traduzam para português. Os serviços empregues são:

- FreeTranslation (http://www.freetranslation.com)
- Systran (http://www.systranbox.com/systran/box)
- E-Translation Server (http://www.linguatec.net/online/ptwebtext/index_en.shtml)
- Amikai (http://www.amikai.com/demo.jsp)

O TrAva apresenta posteriormente ao utilizador as traduções automáticas obtidas a partir desses quatro serviços para que sejam então o alvo de um processo de classificação. A classificação é feita através de um formulário próprio onde o utilizador é convidado a introduzir informação relativa aos problemas de tradução ou tipo de erros que ocorrem nas traduções e, simultaneamente, informação adicional sobre a frase original. Toda esta informação é guardada em base de dados, que é imediatamente consultável por todos os utilizadores.

O TrAva permite aos utilizadores pesquisar a informação da base de dados que é construída colectivamente. Os critérios de pesquisa disponíveis são todos aqueles que foram usados para a classificação das traduções e frases originais (nota: a identidade dos participantes é reservada pelo que não é apresentada nas consultas da base de dados, apesar de ficar armazenada no TrAva). Espera-se que essa compilação colectiva de exemplos de tradução automática, que se encontra acessível publicamente, possa ajudar a melhorar a compreensão dos fenómenos que estão envolvidos na tradução automática do português. Para isso convidamos todos os interessados a visitarem http://www.linguateca.pt/trava/ e a participarem na recolha, após a inscrição no TrAva.

5. Comentário final

Com esta breve apresentação das várias ferramentas, esperamos ter despertado um interesse em utilizá-las. O seu interesse justifica os nossos esforços e qualquer "feedback" que queiram nos mandar será apreciado, porque acreditamos que a pesquisa só pode avançar se houver colaboração e troca de informação e ideais entre pesquisadores.

Parte 2 – Universidade de São Paulo, Brasil

6. O Projeto Lácio-Web

O Lácio-Web (LW) (Aluísio et al 2003, Aluísio et al 2004) é um projeto realizado com parceria entre o Núcleo Interinstitucional de Lingüística Computacional (NILC)[2], localizado no ICMC-USP, o IME-USP e a FFLCH-USP, sob a coordenação de Sandra Maria Aluísio. O objetivo desse projeto é divulgar e disponibilizar gratuitamente na Web: a) vários córpus[3] do português brasileiro escrito contemporâneo, representando bancos de textos adequadamente compilados, catalogados e codificados, em um padrão que possibilita fácil intercâmbio, navegação e análise; e b) ferramentas lingüístico-computacionais, tais como contadores de freqüência, concordanciadores e etiquetadores morfossintáticos treinados em grandes córpus anotados manualmente.

---

2 www.nilc.icmc.usp.br/

3 Nesta parte do artigo, optamos por utilizar a forma aportuguesada córpus, tanto para o singular, quanto para o plural, (a exemplo de lápis) para as palavras latinas *corpus* e *corpora*, respectivamente.

O público-alvo do LW é heterogêneo: de um lado lingüistas, cientistas da computação, lexicógrafos, terminólogos etc. e, de outro, não especialistas em geral. O LW é acessado a partir de um portal (http://www.nilc.icmc.usp.br/lacioweb/), que informa os tipos de córpus, ferramentas, todo o material disponível e forma de contribuir com textos para a continuação do projeto, disponibiliza manuais e artigos relacionados e permite, após cadastramento do usuário, o acesso a seus córpus e ferramentas. O portal pretendeu ser didático ao público leigo em geral e estudantes em particular e, para isso, foram incluídos textos explicativos e de ajuda.

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

A primeira disponibilização pública do projeto ocorreu em 20/01/2004 e a segunda no final do projeto, em 30/06/2004. Parte do material adquirido ainda precisa passar por um processo de três fases para ser disponibilizada: a) compilação-formatação dos textos que vieram da Web e de CD-ROMs, b) nomeação sistemática dos arquivos e c) criação de um cabeçalho para os textos com diversas informações, a saber: bibliográficas comuns – título, autoria, local e data de publicação, fonte, editor, língua; de catalogação – tamanho do arquivo, tipo de amostragem, tipo de autoria, sexo do(s) autor(es); e da tipologia textual em quatro categorias: domínios, gêneros, tipos de texto e meios de distribuição, que serão apresentadas na Seção 6.2. Essa subcategorização detalhada é que permite ao usuário do LW a criação de subcórpus de estudo que atendam a suas pesquisas específicas.

6.1 A Constituição do LW

O Lácio-Web tenta preencher uma lacuna em termos de recursos para pesquisa e suporte à criação de ferramentas de PLN para a língua portuguesa do Brasil. Para tanto, é formado por seis córpus: Lácio-Ref, Mac-Morpho, Lácio-Dev, Par-C, Comp-C e Lácio-Sint, descritos abaixo:

1) **Lácio-Ref:** córpus aberto e de referência composto de textos escritos em português brasileiro, respeitando a norma culta, com 4278 arquivos,

totalizando 8.291.818 ocorrências. É um córpus cru (não anotado com informações morfossintáticas, sintáticas ou de nível mais elevado), mas possui anotações da existência de elementos gráficos e anotação de cabeçalho. A grande maioria dos textos está disponibilizada na íntegra.

2) **Mac-Morpho:** córpus fechado e anotado morfossintaticamente, formado por artigos jornalísticos retirados da Folha de São Paulo, ano 1994, dos cadernos Esporte (ES), Dinheiro (DI), Ciência (FC), Agronomia (AG), Informática (IF), Ilustrada (IL), Mais! (MA), Mundo (MU), Brasil (BR) e Cotidiano (CO). Composto de 1.167.183 ocorrências, o córpus foi etiquetado pelo parser Palavras de Eckhard Bick (http://visl.hum.sdu.dk/), revisado manualmente quanto à anotação morfossintática e serviu de treinamento de três etiquetadores morfossintáticos disponíveis na Web (Aluísio, Pelizzoni et al, 2003). O MAC-MORPHO é disponibilizado para *download* em 2 formatos: 1) adequado para pesquisas lingüísticas com o uso de contadores de freqüência ou concordanciadores, por exemplo; 2) adequado ao treinamento de etiquetadores e que, por ter os polilexicais separados[4], teve o tamanho do córpus alterado para 1.221.468 ocorrências.

3) **Lácio-Dev:** córpus projetado para ser um córpus aberto, isto é, de atualização contínua, e com textos que não foram revisados em relação à norma culta. Esse córpus destina-se à avaliação do desempenho de corretores gramaticais do português brasileiro, isto é, pretende servir de *benchmark* para a tarefa de correção gramatical, bem como para análise de inadequações lingüísticas nos textos.

4) **Par-C:** córpus aberto, paralelo, Português-Inglês, que possui, inicialmente, textos de 1 ano de edições da Revista Pesquisa Fapesp, num total de 646 textos em cada língua. O número total de ocorrências desse córpus é de 893.283.

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

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

confecção de glossários e dicionários especializados e outras pesquisas lingüísticas.

6) **Lácio-Sint** (porção etiquetada do córpus Lácio-Ref): córpus fechado e etiquetado automaticamente com lemas, etiquetas morfossintáticas e sintáticas. Diferente do Mac-Morpho, esse córpus será composto por textos de diversos gêneros e contará com ferramentas desenvolvidas no NILC, tais como o *parser* Curupira (Martins et al, 2002).

No total, o Projeto LW possui 5.708 arquivos, totalizando 10.413.524 ocorrências. Os córpus Lácio-Dev e Lácio-Sint serão disponibilizados futuramente, como fruto de pesquisas de doutorado e mestrado, respectivamente.

6.2 Tipologia Textual do LW

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

Gênero textual: o gênero discrimina o texto pela intenção comunicativa e pelo caráter discursivo, isto é, a comunidade (meio) em que circula e as atividades humanas que o tornam relevante. Convencionamos o uso de um supergênero, chamado Literário (LT), um conjunto de gêneros e um conjunto de subgêneros. Os gêneros e subgêneros são dados abaixo:

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

* Esses gêneros, especialmente, advêm do supergênero Literário.

Tipo textual: considera-se "tipo de texto" o modo específico de estruturação de um texto. Refere-se ao texto visto "de dentro", ou seja, suas partes componentes, seu léxico, sua sintaxe, sua adequação ao tema etc. Trata-se de uma lista em constante atualização e que, no momento, é composta de 39 categorias (e "Outros" – tipos textuais não previstos), por ex.: apostila, manual, parecer, reportagem, súmula, testamento etc.

Domínio: é a "área de conhecimento" que tematiza a principal informação veiculada pelo texto. Temos 3 grandes linhas de domínio, denominadas "domínio geral". A cada uma dessas linhas associam-se subdomínios, denominados "domínios específicos". A divisão em termos de domínio geral apresenta as seguintes vertentes:

a) científica: classifica os textos tematizados pela ciência. Esse grupo é composto por seis áreas do conhecimento: Ciências Agrárias, Ciências Biológicas, Ciências da Saúde, Ciências Exatas e da Terra, Ciências Humanas e Ciências Sociais Aplicadas;

b) **religião e pensamento**: envolve os temas metafísicos, espirituais e teológicos (ex.: livros de bruxaria, de auto-ajuda, etc.).

c) **generalidades**: absorve os textos com temas variados e, de modo geral, inseridos num campo tematizado pelo senso comum (ex.: entretenimento). Inclui, além disso, os textos que abordam, de forma não-analítica, temas considerados pela ciência (ex.: ciência e tecnologia, saúde, esporte, etc.).

Meio de distribuição: seleciona o canal por meio do qual o texto foi divulgado ao seu público-alvo, por ex.: Cd-rom (CR), Diário Oficial (DO), Internet (IN), Jornal (JO), Livro (LI), Tese (TE).

6.3 As Ferramentas do LW

O Projeto Lácio-Web disponibiliza várias ferramentas lingüístico-computacionais como concordanciadores, contadores de freqüência e etiquetadores morfossintáticos, treinados com o córpus do português do Brasil anotado manualmente – o MAC-Morpho – e futuramente pretende disponibilizar extratores de termos e alinhadores de textos paralelos. O objetivo é facilitar o estudo de características lingüísticas do português por pesquisadores da área de lingüística e lingüística computacional, assim como melhorar a qualidade dos sistemas desenvolvidos para o português, tais como, tradutores para o português, sumarizadores automáticos e engenhos de busca especializados no português.

As ferramentas podem ser usadas com o Lácio-Ref, com os subcórpus criados pelo usuário ou ainda com o córpus que o usuário tiver carregado para o

LW. Existem três tipos de pesquisa para montagem de subcórpus: (1) a pesquisa simples que faz a busca de textos baseada em meio de distribuição e gênero; (2) a avançada que considera, primeiramente, meio de distribuição, supergênero e gênero textual. Considera também subgênero para os gêneros Informativo e Prosa e outros dados particulares de cada gênero textual, por exemplo: para o gênero Informativo, o nome do periódico e seção/caderno; para o Científico, o nome do autor e para o supergênero Literário, o nome do autor e da obra; e (3) a pesquisa personalizada que oferece grande parte dos dados de cabeçalho das amostras do Lácio-Ref. Por isso, é possível montar um subcórpus refinado em termos de detalhes da bibliografia e da classificação textual.

6.3.1 Contador de Freqüência Padrão

Calcula a freqüência de todas as palavras do córpus escolhido, informando o número total de arquivos, de ocorrências, de palavras únicas, e a variação do vocabulário por meio do cálculo da razão *type/token*. Apresenta também a tabela de freqüência das palavras. A saída desse contador pode ser salva para leitura posterior.

6.3.2 Contador de Freqüência por Palavra

Dada uma ordem decrescente de freqüência das palavras do córpus escolhido, o Contador calcula a freqüência de uma palavra escolhida, apresentando a freqüência das palavras do contexto anterior e posterior a ela. A saída desse contador pode ser salva para leitura posterior.

6.3.3 Concordanciador para córpus sem anotação

O concordanciador implementado gera uma lista enumerada de todas as ocorrências de uma determinada palavra (ou expressão) escolhida de um córpus sem anotação. As opções para córpus no nosso projeto são, atualmente, o subcórpus de pesquisa gerado a partir do Lácio-Ref, o próprio Lácio-Ref ou o córpus do usuário, desde que sem anotação morfossintática.

6.3.4 Concordanciador para córpus anotado morfossintaticamente

O concordanciador implementado gera uma lista enumerada de todas as ocorrências a partir da seleção de uma palavra com a respectiva etiqueta. A opção para córpus no nosso projeto é, atualmente, o próprio MAC-Morpho e

as etiquetas são as que constam do Manual de Etiquetação do MAC-Morpho, disponível para *download* no portal LW.

6.3.5 Etiquetadores Morfossintáticos

Três etiquetadores disponíveis na Web – MXPOST (Ratnaparkhi, 1996), TreeTagger (Schmid, 1994) e o etiquetador de BRILL (Brill, 1995) – foram treinados com o córpus MAC-MORPHO, que possui 1.221.468 palavras. Esse córpus foi separado em uma parte para treinamento (977.161 palavras), que traz 80% de cada um dos 10 cadernos da *Folha de São Paulo* que compõem o córpus, e uma parte para teste (244.307 palavras) com os 20% restantes de cada caderno. A precisão dos etiquetadores por caderno pode ser vista no portal LW.

7. O Projeto 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 córpus bilíngües de áreas técnicas que envolvam o português brasileiro, está sendo construído o COMET, um Córpus Multilíngüe para Ensino e Tradução.

O COMET (www.fflch.usp.br/dlm/comet) é constituído basicamente de três subcórpus: um Córpus Técnico-Científico, um Córpus de Aprendizes e um Córpus de Traduções.

7.1 O Córpus Técnico-Científico (CORTEC)

O CORTEC constitui um córpus técnico-científico de âmbito geral, mas privilegia quatro grandes áreas, determinadas a partir de questionário que submetemos a diversos tradutores profissionais, via Internet, indagando das áreas mais carentes de material de apoio. As respostas apontaram para: Direito Comercial, Informática, Ortodontia e Meio Ambiente.

A outra parte do córpus está sendo construída com todos os córpus compilados pelos alunos do Curso de Especialização em Tradução (CETRAD) – Inglês e da Pós-Graduação da FFLCH-USP, resultantes de projetos diversos (Tagnin 2003[5]), alguns dos quais serviram para a elaboração de glossários destinados a tradutores (disponíveis em http://www.fflch.usp.br/citrat). Numa pri-

[5] Observe-se aqui o paralelismo com as pesquisas realizadas por Maia, relatadas na outra parte deste artigo.

meira etapa (2001), foram construídos córpus 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, totalizando atualmente 5.463.597 palavras em inglês e 2.928.940 em português.

Em sua grande maioria, os textos constituem córpus comparáveis, ou seja, são de tamanho semelhante (em geral, ao redor de 100.000 palavras), compostos de textos dentro de uma mesma área, seguindo padrões semelhantes, tais como: gênero, tipologia textual, extensão, fonte, data etc., em inglês e português. 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, aspecto praticamente ignorado na grande maioria dos glossários, que em geral se atêm a termos simples – recorremos aqui à Culinária, a título de exemplo – (*pimenta, caçarola*) e compostos (*pimenta-do-reino, pimenta calabresa, panela de pressão*), não registrando unidades de significado mais extensas como *pimenta-do-reino moída na hora*, ou colocações verbais como *untar uma forma, cortar (uma cebola) em rodelas*.

7.1.1 Problemas na construção dos córpus

Ao contrário do que se pode imaginar – que basta "baixar" da Internet textos de um determinado assunto para já se obter um córpus – a construção de um córpus, 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, desde a delimitação da área de estudo e o balanceamento do córpus até a obtenção da permissão de inclusão dos textos, passando pela padronização dos textos e inserção do cabeçalho em cada um deles.

7.1.2 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 córpus da área de Medicina, os seguintes serão agrupados dentro desse domínio:

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

Por se tratar de um córpus 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.

7.1.3 Um córpus paralelo – Revista Pesquisa FAPESP

Dentro do Córpus Técnico-Científico está também sendo construído um córpus paralelo (originais com respectivas traduções) com os textos eletrônicos da Revista Pesquisa da FAPESP, a partir da edição de número 60, do ano 2000, que foram 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 córpus está sendo alinhado no nível da sentença

por um alinhador sentencial desenvolvido no NILC (Caseli e Nunes, 2003). É importante salientar que nesse córpus os textos originais são em português e os traduzidos, em inglês. Além das pesquisas contrastivas léxico-gramaticais de praxe, a diversidade de tipologia textual da revista (reportagens, cartas, notícias, carta do editor, artigos) permitirá estudos também no nível do discurso. Uma parte desse material já está disponível no Par-C do LW.

7.3 O Córpus de Traduções

Esse córpus é constituído de 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 2002). O córpus está sendo aumentado com aproximadamente 25 contos australianos e suas respectivas traduções; b) um conjunto de textos paralelos de literatura brasileira vertida para idiomas estrangeiros. Esse tipo de córpus permite analisar processos e estratégias de tradução, bem como enseja toda sorte de estudos contrastivos, desde morfológicos, sintáticos e lexicais até textuais.

7.4 Corpus de Aprendizes

Esse córpus assemelha-se, em termos de objetivos, ao Lácio-Dev do Lácio-Web, mas é constituído de redações de aprendizes de línguas estrangeiras (alemão, espanhol, francês, inglês e italiano) dos cursos de graduação e extensão da FFCLH/USP e destina-se a pesquisas sobre os problemas mais recorrentes de aprendizagem das respectivas línguas. Permitirá pesquisas tanto horizontais, por exemplo, dificuldades de uma determinada classe ou de aprendizes de certo nível, quanto verticais, ou seja, pesquisas diacrônicas que acompanham o desenvolvimento de um grupo ou indivíduo ao longo de um período de tempo.

Também incluirá um córpus de aprendizes de tradução, onde serão recolhidas as primeiras versões de trabalhos, sem qualquer correção, dos alunos do Curso de Especialização em Tradução do inglês. Com ele pretende-se determinar as dificuldades enfrentadas pelos aprendizes, seja no que concerne a estratégias de tradução de determinados itens, tais como aqueles que indicam aspectos culturais ou dialetais, sejam problemas específicos do próprio vernáculo.

Ambos têm por objetivo último o aperfeiçoamento do ensino nas respectivas áreas.

MAIA, Belinda; SARMENTO, Luís e TAGNIN, Stella E. O. e ALUÍSIO, Sandra Maria. *Idéias que cruzam o oceano.*

7.5 Caracterização do COMET

7.5.1 Os objetivos

Face ao acima exposto, tornam-se claros os objetivos desse córpus no campo da tradução e do ensino. 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 e das línguas estrangeiras, como já vem ocorrendo.

7.5.2 O público-alvo

O público a que se destina abrange desde aprendizes e professores de tradução e das línguas estrangeiras contempladas, 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.

7.5.3 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, além dos textos dos aprendizes.

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 coocorrem 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.

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 estará o aprendiz para produzir uma tradução confiável.

Em suma, o COMET é um córpus multilíngüe destinado ao ensino e à pesquisa de línguas e de tradução. Sua configuração facilita seu uso para a resolução de questões práticas, tais como determinar o uso correto de certo termo, ou a palavra que usualmente co-ocorre com outra, assim como se presta para uma gama extremamente variada de estudos acadêmicos. No âmbito de nossa Universidade, está sendo usado para trabalhos sobre lexicologia, terminologia, construção de córpus, processos e estratégias de tradução, dificuldades dos aprendizes e análises contrastivas. Outros trabalhos envolvem a construção de córpus próprios que, ao término, serão incorporados ao COMET. Dessa forma, estabelece-se uma proveitosa troca acadêmica: o COMET alimenta diversos estudos acadêmicos, enquanto córpus resultantes de outros estudos retroalimentam o COMET. Assim, garante-se o enriquecimento e a constante atualização do córpus.

8. Os pontos de contato

Pelo acima exposto, podemos observar que os construtores de córpus técnicos do Projeto COMET, por exemplo, podem se beneficiar do Corpógrafo para a extração de termos e compilação de glossários. O Projeto COMET, parceiro do NILC na construção do LW, contribuiu e continuará contribuindo com os textos para os quais obtiver autorização de inclusão, tanto em inglês, quanto em português, para aumentar, principalmente, o córpus Par-C. O Lácio-Ref, por ser o único córpus de português brasileiro sistematicamente classificado em gênero, tipo de texto, d omínio e meio de distribuição, que está gratuitamente disponível na Web, complementa recursos semelhantes existentes para a variante européia, permitindo pesquisas contrastivas, em especial nas áreas de especialidade. O objetivo de ambas as partes é disponibilizar gratuitamente na Web recursos computacionais e lingüísticos que possam contribuir para o desenvolvimento e enriquecimento das pesquisas nos dois lados do oceano.

MAIA, Belinda; SARMENTO, Luís e TAGNIN, Stella E. O. e ALUÍSIO, Sandra Maria. *Idéias que cruzam o oceano.*

Referências Bibliográficas

ALUÍSIO, S. M.; PELIZZONI, J. M.; MARCHI, A. R.; OLIVEIRA, L. H.; MANENTI, R.; MARQUIVAFÁVEL, V. (2003). An account of the challenge of tagging a reference corpus of Brazilian Portuguese. In: *PROPOR'2003, 2003, Faro. Lecture Notes on Artificial Intelligence. Proceedings of PROPOR'2003*. Springer Verlag, 2003. v. 1.

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

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

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

BRILL, E. Transformation-based error-driven learning of natural language: A case study in part of speech tagging, *Computational Linguistics 21* (1995), 543-565. Disponível em: http://www.cs.jhu.edu/~brill/.

CASELI, H.M.; NUNES, M.G.V. Sentence Alignment of Brazilian Portuguese and English Parallel Texts. In: *Argentine Symposium on Artificial Intelligence (ASAI 2003)*. Buenos Aires, Argentine, September 2003.

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

FRANKENBERG-GARCIA, Ana & Diana SANTOS. 2003 COMPARA, um corpus paralelo de português e inglês na Web. In: Stella TAGNIN (Org.). *Cadernos de Tradução*, n. 9 – 2002/1, Núcleo de Tradução – NUT, Universidade de Santa Catarina, 61-79.

FRANKENBERG-GARCIA, Ana & Diana SANTOS. 2003. Introducing COMPARA, the Portuguese-English parallel translation corpus. In: Federico Zanettin, Silvia Bernardini & Dominic Stewart (Eds.). *Corpora in Translation Education*, Manchester: St. Jerome Pub. 71-87.

LARANJINHA, A. L. T. *Para um Glossário Bilíngüe Português-Inglês de Termos do Direito Comercial: Colocações Verbais,* dissertação de mestrado, Universidade de São Paulo, 1999.

MAIA, Belinda 2003a Training Translators in Terminology and Information Retrieval using Comparable and Parallel Corpora. In: F. Zanettin, S. Bernardini & D. Stewart *Corpora in Translator Education*, Manchester: St. Jerome Pub. , 43-54.

MAIA, Belinda. 2003b. Ontology, Ontologies, General Language and Specialised Languages. In: *Volume Comemorativo dos 25 anos do CLUP*. Porto: CLUP. 23-39.

MAIA, Belinda 2003c. What are comparable corpora? In: *Proceedings of pre-conference workshop Multilingual Corpora: Linguistic Requirements and Technical perspectives* at Corpus Linguistics 2003, Lancaster U.K., 27-34.

MAIA, Belinda. 2003d. Ensinar como especializar-se. In: *Actas do V Seminário de Tradução Científica e Técnica em Língua Portuguesa*, Lisboa: União Latina.

MAIA, Belinda 2003e. Using Corpora for Terminology Extraction: Pedagogical and Computational Approaches. In: B. Lewandowska-Tomasczczyk, (Ed.). *PALC 2001 – Practical Applications of Language Corpora.*, Lodz Studies in Language. Frankfurt: Peter Lang, 147-164.

MAIA, Belinda. 2003f. Terminology – where to find it, and how to keep it. (Keynote Speaker). In: *Proceedings of III Jornadas sobre la formación del traductor e intérprete*, Universidad Europea de Madrid. CD-ROM.

MAIA, Belinda. 2003g. Do-it-yourself, disposable, specialized mini corpora – where next? Reflections on teaching translation and terminology through corpora. In: Stella Tagnin (Org.). *Cadernos de Tradução*, n. 9 – 2002/1 – Núcleo de Tradução – NUT, Universidade Federal de Santa Catarina, 221-235.

MAIA, Belinda. 2002a. Nothing is inherently boring – reflections on training translators in terminology. In: Maia, B. J. Haller & M. Ulrych (Eds.). 2002, *Training the Language Services Provider for the New Millenium. Proceedings of Encontros III de Tradução da AsTra-FLUP 25-26 Maio de 2001*, 355-64.

MAIA, Belinda. 2002b. The Industrialisation of Translation – will it work? In: *Génesis– Revista Científica do ISAI*, n. 2. Porto: ISAI., 17-26.

MAIA, Belinda. 2002c. Corpora for terminology extraction – the differing perspectives and objectives of researchers, teachers and language services providers. In: The *Proceedings of the Workshop Language Resources for Translation Work and Research* – held in conjunction with LREC 2002, Universidad de Las Canárias, Spain, 25-28.

MAIA, Belinda, 2000. Making corpora – a learning process. In: Bernardini, S. & Zanettin, F. (Eds.). 2000: *I corpora nella didattica della traduzione*. Bologna: CLUEB, 47-56.

MAIA, Belinda 1997. Do-it-yourself corpora... with a little bit of help from your friends! In Barbara Lewandowska-Tomaszczyk and Patrick James Melia (Eds.). *PALC '97 Practical Applications in Language Corpora*. Lodz: Lodz University Press, 403-410.

MAIA, Belinda, & Luís SARMENTO. 2003a. The Pedagogical and Linguistic Research Applications of the GC to Parallel and Comparable Corpora. In: *Proceedings of CP3A 2003: Corpora Paralelos, Aplicações e Algoritmos Associados*. Braga: Universidade do Minho.

MAIA, Belinda and Luís SARMENTO. 2003b. Constructing comparable and parallel corpora for terminology extraction – work in progress. Poster apresentado em Corpus Linguistics 2003, Lancaster U.K. (Vencedor do 1o. prêmio).

MARTINS, R. T.; HASEGAWA, R.; NUNES, M.G.V. Curupira: um parser funcional para o português, *NILC-TR-02-26*, dezembro 2002.

RATNAPARKHI, A. A Maximum Entropy Part-of-Speech Tagger, *Proceedings of the First Empirical Methods in Natural Language Processing Conference* (1996).

SANTOS, Diana, Belinda MAIA & Luís SARMENTO. Gathering empirical data to evaluate MT from English to Portuguese. In: Lambros Kranias, Nicoletta Calzolari, Gregor Thurmair, Yorick Wilks, Eduard Hovy, Gudrun Magnusdottir, Anna Samiotou & Khalid Choukri (Eds.). *Proceedings of LREC 2004 (Workshop on the Amazing Utility of Parallel and Comparable Corpora*) (Lisboa, Portugal, 25 May 2004), 14-17.

SANTOS, Diana. DISPARA, a system for distributing parallel corpora on the Web. In: Nuno Mamede & Elisabete Ranchhod (Eds.). *Portugal for Natural Language Processing* (PorTAL 2002) (Faro, Portugal, 23-26 June 2002), Berlin/Heidelberg: Springer-Verlag. *Lecture Notes in Artificial Intelligence,* 209-218.

SARMENTO, Luís & Belinda MAIA. Gestor de corpora – Um ambiente Web integrado para Lingüística baseada em Corpora. In: José João Almeida (Ed.). *Corpora Paralelos, Aplicações e Algoritmos Associados (CP3A)* (Braga, Junho), Braga: Universidade do Minho, 25-30.

SARMENTO, Luís, Belinda MAIA & Diana SANTOS. The Corpógrafo – a Web-based environment for corpora research. In: Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (Eds.). *Proceedings of LREC 2004* (Lisboa, Portugal, 26-28 May 2004), 449-452.

SCHMID, H. Probabilistic part-of-speech tagging using decision trees, *Proceedings of International Conference on New Methods in Language Processing* (1994), 44-49.

TAGNIN, S. E.O. 2003. Os Corpora: instrumentos de auto-ajuda para o Tradutor. In: Stella Tagnin (Org.). *Cadernos de Tradução*, n. 9 – 2002/1*,* número especial sobre Corpus e Tradução, Universidade Federal de Santa Catarina, Florianópolis: Núcleo de Tradução, 191-219.

TAGNIN, S.E.O. (Org.). 2002. *Lá do Canadá – contos.* São Paulo: Olavobrás.

VARANTOLA, Krista. 2003. Disposable Corpora as Intelligent Tools in Translation. In: Stella Tagnin (Org.).*Cadernos de Tradução*, n. 9 – 2002/1 – Núcleo de Tradução – NUT, Universidade Federal de Santa Catarina, 171-189.