Construção de um corpus paralelo e alinhado português-italiano-português para o domínio literário

Claudia Zavaglia[1], Monique Lopes Ferraresi[2]

[1]Professora assistente doutora no Instituto de Biociências, Letras e Ciências Exatas – Universidade Estadual Paulista (UNESP)
zavaglia@ibilce.unesp.br

[2] Pós-graduanda em Estudos Lingüísticos no Instituto de Biociências, Letras e Ciências Exatas – Universidade Estadual Paulista (UNESP)
moniqueferraresi@yahoo.com.br

Abstract. *This paper describes the elaboration of a parallel corpus and aligned of the literary domain. This corpus is composed of works in Brazilian Portuguese and Italian produced in the 70's until 2004 and it can be used in various applications in Translation Studies and the Natural Language Processing - NLP, as for the Machine Translation (MT), in the Recovery of the Information (RI) and in the Learning of Languages..*

Keywords. *Corpus linguistics; parallel corpus; aligned corpus.*

Resumo. *Este artigo descreve a elaboração de um* corpus *paralelo e alinhado do domínio literário em italiano e português. Este* corpus *é composto de obras em língua portuguesa do Brasil e italiana produzidas dos anos 70 a 2004 e poderá ser utilizado tanto em diversas aplicações em Tradutologia e no Processamento de Línguas Naturais - PLN, como para a Tradução Automática (TA), na Recuperação da Informação (RI) e no Aprendizado de Idiomas.*

Palavras-chave. *Lingüística de corpus; corpus paralelo; corpus alinhado.*

1. Introdução

Com a chegada dos microcomputadores pessoais, nos anos 80, houve uma grande popularização de *corpora* e das ferramentas de processamento automático, fato esse que colaborou, sobremaneira, para o fortalecimento das pesquisas lingüísticas fundamentadas em bases de dados computadorizadas.

Hoje, a Lingüística de *Corpus* (doravante LC) exerce ampla influência na pesquisa lingüística, principalmente na Europa. No Brasil, esse tipo de pesquisa ainda está em estágio inicial, mas não deixa de ocorrer, por exemplo, em centros especializados em Processamento de Linguagem Natural, Lexicografia e Lingüística Computacional.

Desde a mudança do paradigma teórico detectada nos anos 50, quando as teorias racionalistas da linguagem como a Lingüística Gerativa, que sobrepunha-se ao empirismo, deram lugar aos trabalhos baseados em *corpora*, a LC apresenta um crescimento vertiginoso e uma importância imensurável nos atuais estudos lingüísticos. O *corpus* pioneiro de amostragem chamado de Brown (1964) surge na década de 60 e é tido como fator propulsor do desenvolvimento da LC, seguido de outros, como o Banco de Português compilado pela PUC (BP) no projeto *Direct*, que é um *corpus* orgânico já

que está aberto e cujo conteúdo está em constante expansão e renovação (BERBER SARDINHA, 2004).

O uso da tecnologia computacional em tradução vem se tornando cada vez mais freqüente. Há programas de tradução automática, bancos terminológicos, dicionários eletrônicos, bibliotecas informatizadas, além de diversos outros recursos. Nesse cenário, os *corpora* eletrônicos ganham cada vez mais destaque, como recursos de grande utilidade para o tradutor, seja na sua prática profissional ou como pesquisador (BERBER SARDINHA, 2003). Além disso, os *corpora* auxiliam professores de línguas na criação de materiais didáticos, ou tradutores na elaboração de glossários especializados. De fato, Tagnin (2002) narra que:

> Embora o curso não pretendesse enfocar especificamente a linguagem técnica, a maior parte das unidades fraseológicas caracterizava-se como termos técnicos dentro da área investigada. Foi isso que fez com que fosse sugerido aos alunos organizarem essas unidades sob a forma de glossário. Assim, além de construírem um corpus bilíngüe comparável, de proporções bem menores do que inicialmente proposto, cada grupo apresentou um glossário de 50 a 200 termos em cada língua. Os glossários apresentaram os termos equivalentes com exemplos autênticos em ambas as línguas. Não havia definições, pois não pretendia ser um recurso terminológico propriamente dito, isto é, um glossário definitório. Pretendia ser uma fonte de referência para o tradutor, oferecendo-lhe os termos técnicos, seus equivalentes e, acima de tudo, contextos de uso em ambas as línguas. (TAGNIN, 2002, p. 200-201)

A utilização de *corpora* por tradutores funciona, segundo Zavaglia (2004), como um instrumento necessário à prática profissional dos mesmos, possibilitando a consulta e a exploração de equivalências lingüísticas sob diversos aspectos, tais como: o lexical (buscando equivalentes e/ou possíveis traduções para lexias singulares, para expressões idiomáticas e/ou cristalizadas e para fraseologismos); o discursivo (no reconhecimento de enunciados próprios e característicos); o pragmático (no reconhecimento de elementos culturalmente marcados).

É importante ressaltar que a produção de *corpus* destacou-se a partir do momento que os lingüistas "puros" e os lingüistas computacionais confirmaram a capacidade de realizarem pesquisas em *corpora* como um filtro, isto é, exercendo um papel intermediário de suas legitimações, hipóteses e evidências para pesquisas lingüísticas de caráter diverso. Assim, iniciou-se a produção intensa tanto de grandes repertórios textuais nas mais diversas línguas quanto de ferramentas e programas para a sua realização. (ZAVAGLIA, 2004)

Entretanto, ainda é incipiente (senão inexistente) a produção ou elaboração desse tipo de material lingüístico (*corpora*) para o português-italiano e vice-versa, em qualquer domínio, do literário ao técnico-científico. Nesse sentido, acreditamos que a proposta de nossa pesquisa justifica-se plenamente.

2. Objetivos

O presente trabalho objetiva apresentar um protótipo de elaboração de *corpus*, a partir de textos paralelos, ou seja, textos originais acompanhados de sua tradução, nas línguas italiana e portuguesa do Brasil. Esse *corpus* será composto por textos do

domínio literário produzidos nos anos 70, 80, 90 (do século XX) e de 2000 a 2004 (do século XXI) que tenham sido traduzidos de uma língua para a outra.

A partir dos textos selecionados, temos como escopo específico ordená-los paralelamente e alinhá-los sentencialmente. Seu armazenamento dar-se-á na ferramenta computacional para o gerenciamento de grandes bases textuais, *Folio Views*. Almejamos, ademais, propor uma interface Web para esse delineamento de *corpus* que será disponibilizada on-line futuramente.

3. Embasamento teórico

Segundo Berber Sardinha (2000):

> A Lingüística de Corpus ocupa-se da coleta e exploração de corpora, ou conjuntos de dados lingüísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador. (BERBER SARDINHA, 2000, p.2)

Dessa forma, Kennedy (1998) complementa:

> Embora o escopo da Lingüística de Corpus possa ser definido em termos do que as pessoas fazem com corpora, seria um engano assumir que Lingüística de Corpus é somente um meio mais rápido de descrever como a linguagem funciona [...] A análise de um corpus pode revelar, e freqüentemente revela, fatos a respeito de uma língua que nunca se pensou em procurar. (KENNEDY, 1998, p.37)

Na literatura da Lingüística de *Corpus* (LC) há várias definições de *corpus*, propostas por diversos pesquisadores. Para McEnery e Wilson (1996), um *corpus* pode ser entendido como sendo um conjunto de dados que contém mais de um texto. Por sua vez, Sinclair (1991, p. 171) propõe que *corpus* seja "Uma coletânea de textos naturais, escolhidos para caracterizar um estado ou variedade de linguagem". No entanto, essa definição não explicita o propósito da criação de um *corpus*. Dessa forma, a definição a seguir completa a anterior dizendo que "[Corpus é] um corpo de linguagem natural (autêntica) que pode ser usado como base para pesquisa lingüística". (SINCLAIR, 1991, p.10). Em consonância, Sanchez (1995) propõe uma definição que apresenta as características principais que um *corpus* deve conter:

> "Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise". (SANCHEZ, 1995, p.11)

Assim sendo, definimos o item lexical *corpus* como sendo um conjunto de textos produzidos em uma determinada língua natural que caracteriza e reflete o uso sincrônico dessa língua em uma comunidade lingüística, podendo variar entre o registro falado e o escrito. E ainda, um *corpus* pode ser entendido como sendo um arquivo que funciona como uma espécie de depósito de textos ainda não organizados, que podem ser estruturados numa biblioteca eletrônica e separados na mesma, de acordo com interesses específicos.

Segundo Pinheiro (2003), muitos *corpora* têm sido construídos para o português, dentre eles os *corpora* NURC-RJ e NURC-SP, o *corpus* da UNESP de Araraquara, o *corpus* do NILC (Núcleo Interinstitucional de Lingüística Computacional – USP/São Carlos), e mais recentemente o Lácio-Web (LW)[i], cuja proposta é a divulgação e a disponibilização livre na Web de diversos *corpora* do português brasileiro escrito e contemporâneo, representando bancos de textos compilados, catalogados e codificados de forma adequada e em um padrão que possibilite fácil intercâmbio, navegação e análise. Além disso, o LW objetiva oferecer ao público ferramentas lingüístico-computacionais, tais como contadores de freqüência, concordanciadores e etiquetadores morfossintáticos livremente na Web.

Existem pontos importantes a serem observados na construção de um *corpus*. Os dados usados no trabalho de criação de um *corpus* precisam ser necessariamente autênticos. Um *corpus* serve como objeto para estudos lingüísticos. Seu conteúdo é criteriosamente selecionado, a fim de seguir as pressupostas condições de *naturalidade* e *autenticidade* e respeitar as normas estabelecidas pelos criadores do *corpus*.

É imprescindível que os textos do *corpus* encontrem-se legíveis em sua formatação, uma vez que informações ilegíveis podem dificultar o uso por pesquisadores. Com relação à representatividade, sabemos que um *corpus* representa uma determinada língua ou variedade lingüística. Essa representatividade é proporcional à sua extensão, ou seja, quanto maior ele for, mais representativo será.

Segundo Berber Sardinha, (2004), existem alguns pré-requisitos para formar um *corpus* computadorizado aqui resumidos:

(i) Os textos não podem de maneira alguma terem sido criados numa linguagem artificial, pois eles devem ser autênticos e em linguagem natural.

(ii) Em segundo lugar, os textos utilizados devem ser escritos por falantes nativos (caso contrário, o *corpus* é considerado um *corpus* de aprendizes). (BERBER SARDINHA, 2004 p.19-20)

3.1. Tipologia do corpus

Para elaboração de um *corpus*, alguns critérios são fundamentais, dentre eles: representatividade, extensão e tipologia. No que diz respeito à tipologia, elencamos como tipos de *corpus* o **falado** e o **escrito**. Um *corpus* falado possui falas transcritas, enquanto num *corpus* escrito constam textos escritos.

Em relação à temporalidade, um *corpus* pode ser **sincrônico**, quando é datado de um só período; **diacrônico**, quando se refere a vários períodos de tempo; **contemporâneo**, em se tratando de um tempo atual, e finalmente, temos um *corpus* **histórico**, quando se trata de textos que refletem um tempo passado. Existe, ainda, uma classificação que pressupõe um ***corpus* de amostragem**, caso o mesmo indique uma amostra finita de uma linguagem; **monitor**, quando reflete o estado atual do banco de dados; e é um ***corpus* dinâmico** ou **orgânico** aquele que cresce e decresce. Opondo-se ao dinâmico, temos o ***corpus* estático**, ou seja, que não possui variabilidade (nem cresce e nem decresce) e por fim, um *corpus* distribuído em quantidades semelhantes que é chamado de **equilibrado** em virtude da divisão homogênea de seus dados.

Cada *corpus* em potencial possui um tipo de conteúdo. Um ***corpus* especializado** utiliza textos específicos em seu banco de dados. Quando o *corpus* é **regional** ou **dialetal** trata da análise de variações sociolingüísticas específicas. Há ainda

um outro tipo de *corpus* chamado **multilingüe** para designar bases textuais que incluam idiomas diferentes.

Quando chamamos um *corpus* **de aprendiz** significa que a linguagem do *corpus* é proveniente de falantes não nativos de uma dada língua; em contrapartida, um *corpus* pode ser de língua **nativa** se, ao contrário do primeiro, se relacionar com textos cujos autores são necessariamente falantes nativos da língua.

Em se tratando de finalidade, temos:

a) ***corpus* de estudo**: usado para descrição dos textos selecionados;
b) ***corpus* de referência**: utilizado como contraste com o *corpus* de estudo;
c) ***corpus* de treinamento ou teste**: criado para desenvolver aplicações e/ou ferramentas de análise.

Quanto à disposição interna, um *corpus* pode ser **paralelo**, quando possui textos comparáveis; e **alinhado**, quando as traduções aparecem abaixo de cada linha do original.

A proposta de *corpus* deste projeto aglutina a disposição paralela juntamente à alinhada.

Segundo Caseli e Nunes (2004):

> Textos paralelos, segundo a terminologia estabelecida pela comunidade de lingüística computacional, são textos acompanhados de sua tradução em uma ou várias línguas. São considerados distintos dos textos sobre um mesmo tópico, escritos em línguas diferentes, mas que não são, necessariamente, traduções mútuas: os textos comparáveis. (CASELI e NUNES, 2004, p.581)

3.2. Representatividade do *corpus*

Já em relação à representatividade e à extensão, parte-se da premissa de que todo *corpus* possui uma função representativa, e, para ser representativo, o conjunto de textos deve ser o maior possível, ou seja, deve possuir uma dada extensão de um número determinado de palavras e de textos. Segundo Sinclair (1991), o modo principal ou 'salvaguarda', por meio da qual é possível garantir maior representatividade de um *corpus* é através do aumento da sua extensão. Assim, um *corpus* será mais representativo quanto maior ele for, devido ao fato de conter mais instâncias de traços lingüísticos raros.

Nesse sentido, Biderman (2001) revela:

> No desenho do *corpus* é necessário que haja uma proporção equilibrada dos diferentes tipos de textos e/ou de temas nele incluídos. É também importante que o *corpus* seja representativo dos diferentes gêneros e variedades dos usos lingüísticos, ou seja, impõe-se a representatividade dos diferentes níveis de linguagem para assegurar a inclusão de todos os aspectos do idioma. Só assim o *corpus* pode representar, em miniatura, o universo multifacetado da língua. Quando se projeta um *corpus* visa-se extrair de sua observação generalizações sobre a língua. Portanto, não se pode atribuir um peso excessivo a um gênero ou a outro. (BIDERMAN, 2001, p.79)

Como não existem critérios universais para determinação de uma representatividade que atenda as noções gerais de *corpus*, a questão da representatividade continua com suas controvérsias.

3.3. Extensão do *corpus*

Para se detectar a extensão de um *corpus*, existem, atualmente, diversas abordagens, entre elas:

(i) **Abordagem impressionística**: baseia-se em constatações advindas da criação ou exploração dos *corpora* para determinar a sua extensão;
(ii) **Abordagem histórica**: é a que parte da monitoração dos *corpora* usados pela comunidade;
(iii) **Abordagem estatística**: utiliza a estatística para determinar a extensão do *corpus*.

4. Metodologia e Desenvolvimento

Com base em quatro parâmetros propostos por Berber Sardinha (2000), a elaboração da nossa proposta de *corpus* segue as seguintes etapas:

(1) Seleção de textos somente em língua natural;

(2) A tipologia de textos apresenta as seguintes características: (i) autenticidade, ou seja, originais produzidos por falantes nativos do italiano e do português, dependendo da direção do *corpus* (italiano-português ou português-italiano) e (ii) não-autenticidade, ou seja, traduções produzidas por falantes não nativos, ou melhor, por aprendizes do italiano e do português, dependendo da direção do *corpus* (italiano-português ou português-italiano);

(3) A composição dos textos abrange os anos 70, 80, 90 (do século XX) e o período de 2000 a 2004 (do século XXI) que tenham sido traduzidos de uma língua para a outra, cuja língua portuguesa seja a vertente brasileira;.

(4) A representatividade do *corpus* pretende ser alcançada na medida em que propomos uma sistematização de textos nessas duas línguas como um protótipo, ou seja, um modelo que possa vir a ser utilizado e ampliado futuramente. Nesse sentido, ele será representativo para a finalidade para a qual pretendemos.

A ordenação paralela dos textos considera o alinhamento sentencial que determina as correspondências entre as sentenças do texto original e de sua tradução. Para tanto, os textos deverão ser digitalizados e itemizados. No momento do alinhamento, as sentenças são etiquetadas para demarcarmos a qual parte do original corresponde àquela parte da tradução. Todos esses dados serão armazenados na ferramenta computacional mencionada anteriormente, caracterizando-se, esta, como uma importante etapa metodológica do trabalho.

Vejamos uma exemplificação do layout da presente proposta de ordenação de *corpus* do domínio literário na direção italiano-português:

Direção: italiano > português	
Già da più di tre ore erano lì; faceva caldo; era un pomeriggio di prima estate, un po' coperto, nuvoloso; nelle armature si bolliva come in pentole tenute a fuoco lento. <so>[ii]	*Encontravam-se ali havia mais de três horas; fazia calor, era uma tarde de começo de verão, meio encoberta, nebulosa; quem usava armadura fervia como se estivesse em panelas em fogo baixo.*<st>[iii]

As etiquetas <so> e <st> indicam, respectivamente, o final da sentença do texto original (no caso, em língua italiana) e o final da sentença do texto traduzido (no caso, em língua portuguesa). Além dessas, haverá outras etiquetas que indicarão os autores, as obras, os capítulos com as quais trabalharemos.

Baseando-se no *corpus* COMPARA[iv] optamos por um alinhamento sentencial que se baseia nas frases dos textos originais. Nos casos em que não houver correspondência direta na divisão frásica do texto de partida e do texto de chegada, optaremos por dividir ou juntar as frases da tradução de acordo com o formato do original. Dessa forma, poderemos comparar "original com tradução", além de "duas ou mais traduções" de um único original. As frases do original que forem deliberadamente ou por um descuido omitidas na tradução encontrar-se-ão alinhadas com unidades em branco. Já as frases acrescentadas à tradução sem correspondência no texto original, encontrar-se-ão anexadas à unidade de alinhamento precedente mais próxima. (FRANKENBERG-GARCIA & SANTOS, 2002)

4.1. A ESTRUTURA DO CORPUS

A começar do layout do *corpus* do projeto COMPARA[v], propomos a elaboração do **CORPIL** – *Corpus Português-Italiano Literário*, ou seja, um conjunto aberto de textos em português do Brasil alinhado com as suas traduções para o italiano, e de textos em italiano alinhados com as suas traduções para português.

Será possível realizar uma *Busca Simples* para a pesquisa ou então uma *Busca Ampliada*. Na *Busca Simples*, será possível pesquisar palavras e expressões em todos os textos do *corpus* e os resultados das suas buscas são apresentados em forma de concordâncias paralelas alinhadas sentencialmente, ou seja, sentença a sentença. Na *Busca Ampliada*, será possível obter pesquisas mais refinadas e escolher o tipo de *corpus* que se deseja trabalhar ou o conjunto deles para que se obtenha diferentes resultados.

Vejamos os exemplos do layout proposto para o *corpus*:

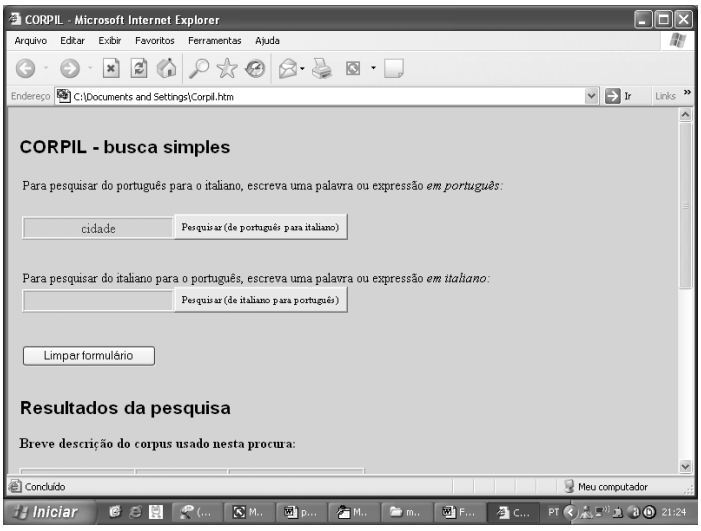

Figura (1) – Proposta de layout para o Corpil – Busca simples.

Figura (2) – Proposta de layout para o Corpil – Resultados da busca simples.

Vejamos os resultados para uma busca ampliada:

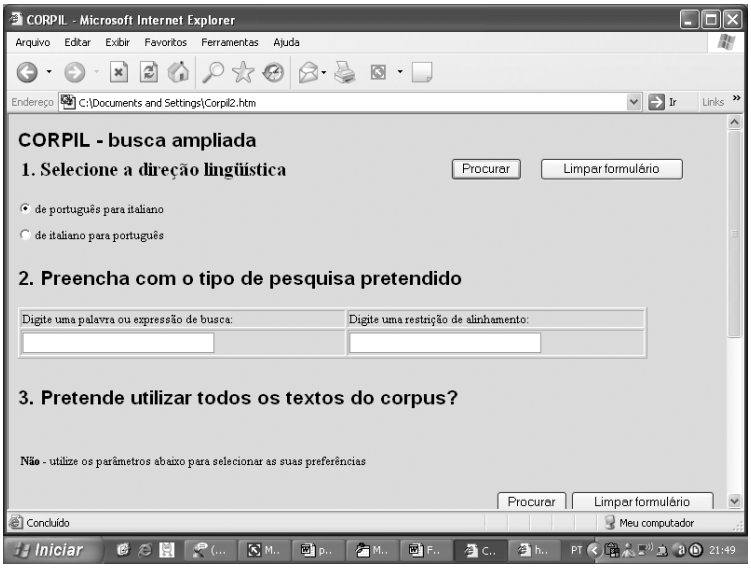

Figura (3) – Proposta de layout para o Corpil – Busca ampliada.

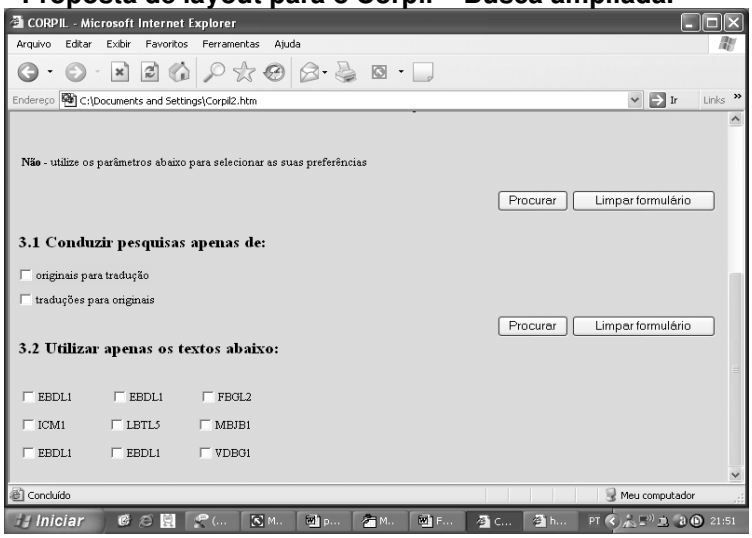

Figura (4) – Proposta de layout para o Corpil – Resultados da busca ampliada.

5. Considerações Finais

Até o presente momento, implementamos o protótipo numa versão preliminar e o nosso corpus conta com cerca de 100.000 ocorrências a partir de cinco obras originais e cinco traduções, a saber: CALVINO, I. Il Cavaliere Inesistente (1993) / O cavaleiro inexistente (); CALVINO, I. Marcovaldo ovvero Le stagione in città (1993)/ Marcovaldo ou As estações na cidade (1994); CALVINO, I. Palomar (1994) / Palomar (1994); GINZBURG, N. La Strada che va in Città (1998) / O caminho que leva à cidade (1998); PALAZZESCHI, A. Sorelle Materassi (1990) / Irmãs Materassi (1993). Em seguida, pretendemos refinar o protótipo em sua versão computacional, além de prosseguirmos com o armazenamento de outras obras literárias.

---

[i] http://www.nilc.icmc.usp.br/lacioweb/index.htm

[ii] Trecho extraído de: CALVINO, ITALO**. Marcovaldo ovvero Le stagioni in città**. Arnaldo Mondadori Editore S.p.A.: Milano, 1993.

[iii] Trecho extraído de: CALVINO, ITALO. **Marcovaldo ou As estações na cidade.** Tradução de Nilson Moulin. Companhia das Letras: São Paulo, 1994.

[iv] Corpus paralelo bi-direcional e extensível que tem como base textos de ficção originais e traduções em português e inglês. Disponível gratuitamente em: http:// www.portugues.mct.pt/COMPARA/BemVindo.html.

[v] A título de apresentação do projeto baseamo-nos na implementação do COMPARA; futuramente, pretendemos apresentar inovações, bem como diferenciais do nosso trabalho.

6. Referências Bibliográficas

BERBER SARDINHA, T. Lingüística de Corpus: histórico e problemática. *DELTA*, 2000, vol.16, no.2, p.323-367. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005&lng=en&nrm=iso Acesso em: 14/12/2004

_____. *Lingüística de Corpus.* Manole: Barueri, 2004.

_____. Uso de corpora na formação de tradutores. *DELTA*, vol.19, no.spe, p.43-70, 2003. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502003000300005&lng=en&nrm=iso Acesso em: 14/12/2004

BIDERMAN, M. T. C. *Teoria Lingüística:* teoria lexical e lingüística computacional. 2º edição. São Paulo: Martins Fontes, 2001.

CASELI, H. M.; NUNES, M. G. V. Corpus paralelo e corpus alinhado**:** propriedades e aplicações. *Estudos Lingüísticos*, 2004. Disponível em: www.nilc.icmc.usp.br/nilc/download/GELCaseli04.pdf Acesso em: 10/12/2004

FRANKENBERG-GARCIA, A; SANTOS, D. COMPARA, um corpus paralelo de português e inglês na Web. In: Tagnin, S. E. O. (org.). *Cadernos de Tradução:* Corpora e Tradução. Florianópolis: NUT, 2002, v. 1, n. 9, p. 61-79.

KENNEDY, G. *An introduction to corpus linguistics***.** Nova York, Longman, 1998.

McENERY, T.; WILSON, A. *Corpus linguistics.* Edinburgh, Edinburgh University Press, 1996.

PINHEIRO, G. M. Projeto Lacio-Web: Panorama Geral e Questões de Design na Criação de um Corpus de Referência do Português do Brasil. *Estudos Lingüísticos*, 2003.

SANCHEZ, A. Definición e historia de los corpus. In: SANCHEZ, A. et al. (org.). *CUMBRE:* corpus linguistico de Español contemporaneo. Madrid: SGEL, 1995, p.7-24.

SINCLAIR, J. *Corpus, Concordance, Collocation.* Oxford: Oxford, University Press, 1991.

TAGNIN, S. E. O. Os Corpora: instrumentos de auto-ajuda para o tradutor. In: Tagnin, S. E. O. (Org.). *Cadernos de Tradução:* Corpora e Tradução. Florianópolis: NUT, 2002, v. 1, n. 9, p. 191-218.

ZAVAGLIA, C. Córpus Lingüístico Paralelo Português-Italiano para a Tradução Juramentada (CLiPPI-Trad_Jura): elaboração e aplicação. In: CONGRESSO IBERO-AMERICANO DE TRADUÇÃO E INTERPRETAÇÃO – III. *Anais*... São Paulo: UNIBERO, 2004. CD-ROM.