Pandaemonium germanicum 6/2002, 237-251

Lingüística de *corpus*: conceito, noções gerais e aplicação

Eurides Avance de Souza* & Iris Kurz Gatti**

Abstract: This paper presents an overview of *Corpus* Linguistics and some possibilities of studies with *corpora*. It gives suggestions on how to build a *corpus* and shows the application of *Corpus* Linguistics in different areas of linguistic research.

Keywords: *Corpus* Linguistics; linguistic research.

Zusammenfassung: Dieser Aufsatz präsentiert einen Überblick über die Korpuslinguisvik und stellt einige Untersuchungsmöglichkeiten mit *Korpora* dar. Er gibt Ratschläge zum Aufbau eines *Korpus* und berichtet über die Verwendung der Korpuslinguistik in verschiedenen Bereichen der linguistischen Forschung.

Stichwörter: Korpuslinguistik; linguistische Forschung.

Palavras-chave: Lingüística do corpus; pesquisa lingüística.

1. Introdução

É cada vez mais freqüente o uso de *corpora* em pesquisas sobre língua, porque eles constituem uma fonte de consulta segura para o pesquisador, oferecendo a ele exemplos lingüísticos autênticos. Aarts afirma:

* Mestranda na área de Língua Alemã na Universidade de São Paulo.
** Mestranda na área de Língua Alemã na Universidade de São Paulo.

> "... the *corpus* in its raw form is for the *corpus* linguist the testbed for his hypotheses about the language..." (AARTS 1991: 45)

Portanto, o pesquisador deve sempre estar atento a suas intuições sobre a língua, sem deixar de lado a observação das questões pesquisadas em uma fonte de dados lingüísticos autênticos, a fim de confirmar aquilo que intui.

O objetivo deste artigo é fornecer um panorama da Lingüística de *Corpus* e das possibilidades de estudos com *corpora*. São apresentados o histórico desta metodologia, definições e tipos de *corpora*, sugestões para compilação de um *corpus*, e aplicação da Lingüística de *Corpus* em diferentes áreas de pesquisa lingüística.

2. O que é um *corpus*?

Atualmente, quando se fala em *corpus*, pensa-se em uma coleção de textos armazenada como um banco de dados eletrônico. Segundo a definição de SANCHEZ (1995), um *corpus* é

> "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 descrição e análise." (SANCHEZ 1995: 8-9)

3. A Lingüística de *Corpus*

A Lingüística de *Corpus* não é uma área de investigação propriamente dita – como a sintaxe, a semântica, a sociolingüística – e sim uma metodologia que pode auxiliar qualquer área de estudos lingüísticos, pois explora a linguagem por meio de evidências empíricas, extraídas por computador.

Comparado a seus antecessores da época pré-computador, um *corpus*, como é compreendido atualmente, oferece muitas possibilidades de pesquisa e facilidades para o trabalho do pesquisador, pois

- é de fácil acesso, seja por meio de CD-ROMs ou Internet;
- possibilita a realização de buscas, extração e classificação de ocorrências por meio de programas de computador;
- possibilita a manipulação de grande quantidade de dados;
- permite estudos de freqüência que podem dar informações importantes, tanto em relação aos dados encontrados, quanto aos não encontrados;
- facilita a obtenção de resultados quantitativos/ estatísticos;
- possibilita a elaboração de concordâncias (listas de ocorrências da palavra ou da expressão procurada, com um pequeno contexto anterior e posterior) por meio de programas de computador;
- permite verificar se determinados fatos se repetem, confirmando intuições e levando a conclusões;
- facilita chegar a generalizações, que podem servir de base para regras gramaticais;
- permite ao não nativo o acesso a dados autênticos;
- permite a elaboração de material de ensino (cf. TAGNIN 2001).

4. Histórico da Lingüística de *Corpus*

Desde o século XIX já se trabalhava com *corpora*, mas havia algumas dificuldades, já que eles não eram informatizados e seu processamento era humano, lento, não confiável e caro.

Na década de 80, com o surgimento dos microcomputadores pessoais, houve a popularização de *corpora* e de ferramentas de processamento, o que ajudou no fortalecimento da pesquisa lingüística baseada em *corpus*.

O histórico da Lingüística de *Corpus* tem uma estreita relação com a disponibilidade de *corpora* eletrônicos. Os primeiros *corpora* compilados foram em língua inglesa, mas atualmente há *corpora* compilados em várias línguas.

Em alemão há o *corpus* do *IDS – Institut für deutsche Sprache* (Mannheim), que, além de diversificado (contém textos jornalísticos, literários etc.), possui ferramentas de busca, elaboração de concordâncias e cálculos estatísticos.

Na UNESP de Araraquara há uma equipe de pesquisadores compilando o *Corpus de Português Contemporâneo*, que atualmente tem 11 milhões de palavras e é considerado o maior *corpus* de língua portuguesa em forma eletrônica.

Na Universidade de São Paulo está sendo desenvolvido o projeto *COMET – Um Corpus Multilíngüe para Ensino e Tradução*, que pretende compilar um *corpus* multilíngüe, tendo o português como língua central, e construir ferramentas de busca e análise de *corpus*. Trata-se de um projeto interdepartamental, reunindo o Departamento de Letras Modernas e o Departamento de Letras Clássicas e Vernáculas, contando com o apoio do Departamento de Ciência da Computação do Instituto de Matemática e Estatística da Universidade de São Paulo (cf. TAGNIN 2001).

5. Tipos de *corpora*

Os *corpora* podem ser compostos por textos de diferentes áreas, dependendo do objetivo ao qual se propõem. Podem conter textos de língua falada ou escrita – atuais ou antigos (para pesquisas de cunho histórico) –, textos jornalísticos, literários ou de alguma área específica (direito, medicina, informática), de linguagem de aprendizes de língua (para estudos sobre aquisição de linguagem, análise de erros) etc.

Tais *corpora* podem ser:

- Monolíngües, bilíngües ou multilíngües;
- anotados (contendo anotação/ classificação morfossintática das palavras dos textos)[1] ou não-anotados;

[1] Outro recurso fornecido pela Lingüística de *Corpus* para auxiliar o trabalho de quem faz pesquisa sobre língua são os anotadores de *corpora*. Esses instrumentos fazem anotações morfossintáticas nos textos, possibilitando buscas pela classe da palavra ou pela sua função na oração.
Um grupo do Instituto de Língua e Comunicação da Universidade do Sul da Dinamarca desenvolve um projeto, sob liderança do Prof. Dr. Eckhard Bick, que disponibiliza um anotador de *corpora* em seu *site* na Internet (http://visl.hum.sdv.dk). Tal anotador pode ser aplicado a textos nas seguintes línguas: alemão, dinamarquês, espanhol, esperanto, francês, inglês, italiano e português.

- paralelos (contendo textos de uma determinada língua e traduções destes textos em uma ou mais línguas) ou comparáveis (contendo originais de determinados tipos de texto em duas ou mais línguas);
- sincrônicos ou diacrônicos;
- abertos (aos quais continuamente são acrescentados textos) ou fechados.

6. Sugestões para compilação de um *corpus*

Muitos pesquisadores compilam seus próprios *corpora* para pesquisa, utilizando textos de revistas, jornais, obras literárias, redações de aprendizes de língua materna ou estrangeira, de acordo com o objetivo da pesquisa.

Diversos jornais e revistas disponibilizam suas edições em CD-ROMs e na Internet. Embora esses textos não constituam *corpora*, eles podem representar a matéria-prima para a construção de *corpora*.

Para compilar seu *corpus* o pesquisador precisa

- delimitar seu tema ou área de pesquisa;
- escolher fontes diversificadas de extração dos textos (dependendo do objetivo do *corpus*);
- coletar os textos com referência, para que seja possível recuperar sua origem, se for necessário;
- organizar os textos de maneira que seja possível identificá-los, ou seja, "etiquetar" os textos, colocando nessas "etiquetas" informações, por exemplo, sobre tipo de texto, fonte, tema, produtor do texto, data etc. (dependendo do objetivo do *corpus*).

A extensão do *corpus* pode variar, de acordo com seu objetivo. Um *corpus* de cerca de 200 mil palavras (para cada língua, no caso de um *corpus* bilíngüe ou multilíngüe) é classificado como pequeno-médio (cf. SARDINHA 1999) e pode ser considerado suficiente para uma pesquisa individual. Obviamente, se o pesquisador não encontrar dados suficientes para análise, seu *corpus* deve ser ampliado.

7. Aplicação da Lingüística de *Corpus*

Face ao rápido e contínuo desenvolvimento tecnológico dos tempos atuais, a Lingüística de *Corpus* está se firmando como uma metodologia de pesquisa eficaz e vantajosa e tem sido aplicada com sucesso em várias áreas.

Assim, seu uso está-se disseminando nas áreas de ensino e aprendizagem de língua, na área de pesquisa em lexicologia e lexicografia (bem como nas subáreas de fraseologia e terminologia/terminografia), nas pesquisas de gramática, estudos contrastivos de língua, tradução, estudos de literatura e estudos de prosódia semântica.

7.1. Ensino e aprendizagem de língua

O ensino de língua estrangeira muito se beneficia com a utilização de *corpora* como instrumento de trabalho. Além de colocar à disposição do professor exemplos autênticos e concretos da língua em uso, proporciona também ao aluno o contato direto e imediato com a língua.

A aprendizagem é um processo ativo, ou seja, que depende da disposição e da atuação do aluno para que se concretize. Ora, a exploração de um *corpus* por parte do aluno irá permitir-lhe, de forma independente, sanar variadas dúvidas, seja em relação à escolha de um vocábulo a ser utilizado em uma redação, seja em relação ao padrão gramatical a ser aplicado quando da utilização de um certo vocábulo (dados referentes à valência, combinações lexicais etc.), além de muitas outras possibilidades. Certamente, esse trabalho ativo de exploração, pesquisa, verificação e constatação irá contribuir sobremaneira para o aprendizado do aluno, espelhando a materialização de consagradas teorias sobre a aprendizagem.

Um exemplo de confusão muito freqüente entre sinônimos, cometida por alunos de língua alemã, diz respeito ao uso indistinto dos verbos *machen* e *tun*. Apenas com o auxílio dos dicionários ou das regras gramaticais não é possível elucidar a diferenciação entre ambos. A possibilidade de recorrer a um *corpus*, para se averiguar as diferenças básicas da utilização dos mesmos, torna-se preciosa.

Em um levantamento que fizemos em *corpora* de língua escrita do IDS, pesquisando ambos os verbos, obtivemos os seguintes resultados:

- ocorrências do verbo *machen* = 171.298
- ocorrências do verbo *tun* = 72.564

Apenas este dado já é bastante significativo, pois que a freqüência é um atributo importante das palavras. Nota-se que *machen* é mais do que duas vezes mais freqüente do que *tun*.

A análise das concordâncias elaboradas para cada um dos verbos demonstrou que *machen* muitas vezes vem acompanhado de um substantivo, com o qual forma uma combinação usual e fixa. Encontramos, portanto, combinações como *Platz machen, Vorwurf machen, Schule machen, Vorschlag machen, Spass machen*. Tais combinações são denominadas, na terminologia da fraseologia, de "colocações", e se caracterizam como ligações convencionais de palavras, geralmente sem motivação semântica clara. Já em relação ao verbo *tun* não se observa esse fenômeno lexical. Muitas outras distinções podem ser extraídas da análise das ocorrências de ambos os verbos. Entretanto, devido à limitação de espaço, não iremos nos aprofundar na análise dos mesmos, restringindo-nos a recomendar uma visita ao *site* do IDS.

Para o professor não nativo, o *corpus* também se consubstancia como uma fonte de pesquisa imprescindível, fornecendo-lhe dados sobre os mais variados aspectos da língua, sobretudo aqueles ligados à convencionalidade e fraseologia, geralmente não dominados por um falante não nativo. É o caso de colocações verbais do tipo, *fazer um bolo* (*einen Kuchen backen*), *passar numa prova/exame* (*eine Prüfung bestehen*), *dobrar a esquina* (*um die Ecke biegen*), *suprir as necessidades* (*den Bedarf decken*), *traçar um plano* (*einen Plan entwerfen*), *tomar a palavra* (*das Wort ergreifen*), *tomar coragem* (*Mut fassen*), muito comuns na linguagem do cotidiano. A simples consulta ao dicionário (seja pelo verbo ou pelo substantivo) não é suficiente para obter a combinatória usual em alemão, o que somente pode ser obtido através de buscas em *corpora*.

Além disso, o *corpus* pode ser usado pelo professor de línguas na preparação de aulas, como material ilustrativo de determinados pontos da gramática e, ainda, como material de apoio para as aulas em que for requisitar a produção de textos no idioma estrangeiro. Segundo Salkie (1997), o *corpus* ajudará os alunos a produzirem sentenças naturais na L2, em lugar de usar estruturas da L1 com o vocabulário da L2.

7.2 Lexicologia e lexicografia

Os estudos de lexicologia podem encontrar na utilização de *corpora* um subsídio essencial para seu desenvolvimento. Isso porque o estudo do léxico de uma língua, sobretudo no que diz respeito ao estudo dos sinônimos, está total-

mente baseado em aspectos concretos de seu uso, seu contexto. Aliás, a própria definição de *sinonímia* oferecida pelo dicionário de semiótica de GREIMAS (1979), bem como a preconizada por GECKELER (1984) e GENOUVRIER (1974), além de outros, estabelece que para haver a sinonímia é preciso que as palavras possam ser comutáveis em qualquer contexto. O acesso ao contexto, fornecido pelo *corpus*, torna possível averiguar as diferenças de uso de palavras tidas como sinônimas, permitindo verificar as situações comunicativas em que são utilizadas.

Em *corpora* particular que possuímos com textos de contratos sociais e estatutos de sociedades anônimas, detectamos a ocorrência de duas colocações, cujos contextos demonstram serem sinônimas. As colocações são: *Handlungen vornehmen* e *Geschäfte tätigen*. Vejamos seus contextos:

(1) "*Die Gesellschaft kann sämtliche* **Geschäfte tätigen**, *welche sie zur Erreichung des Gesellschaftszweck förderlich oder erleichternd erachtet.*"

(2) "*Die Gesellschaft kann alle* **Handlungen vornehmen**, *die mitellbaren oder unmitellbaren Bezug auf ihren Geschäftszweck haben oder für dessen Verwirklichung nützlich sind.*"

O contexto demonstra que as colocações são utilizadas para o mesmo sentido, geralmente formalizado em português através da colocação *praticar atos*. Temos, assim, as respectivas traduções:

(1) A Sociedade poderá **praticar todos os atos** que julgar úteis ou aptos a facilitar o alcance do objeto social.

(2) A Sociedade poderá **praticar todos os atos** que, direta ou indiretamente, tenham relação com seu objeto social ou que sejam úteis para sua consecução.

Depreende-se, pois, que a averiguação de relações léxicas entre vocábulos ou expressões é um dado apreensível em contexto, bem como a configuração exata do sentido de certo vocábulo.

Um estudo minucioso sobre o assunto foi desenvolvido por BIBER (1998), em capítulo dedicado à análise de sinônimos próximos, do idioma inglês, como *little* e *small*, *begin* e *start*. Em tal estudo, constatou que *small* está mais associado à função predicativa em detrimento da função atributiva, e que, quando usado na função atributiva, geralmente co-ocorre com substantivos que indicam quantidade. Já *little*, assim como *big*, tem uma tendência a co-ocorrer com coisas concretas, animadas. O verbo *start* é utilizado como verbo intransitivo com mais freqüência do que o verbo *begin*. Este último, bastante utilizado como verbo transitivo, rege um número de sentenças precedidas de *to* maior do que o *start*.

Assim, segundo BIBER (1998), as pesquisas lexicográficas têm sido incrementadas por meio do uso de técnicas baseadas em *corpus*, as quais estudam os modos como as palavras são usadas, considerando-se, por exemplo, o quão comuns elas são, o quão comum são os diferentes sentidos que podem apresentar, se estão sistematicamente associadas a outras, se estão sistematicamente associadas a registros particulares ou dialetos.

Também para a compilação de dicionários tem se lançado mão do auxílio de *corpora*, seja para elaborar dicionários de língua geral, seja para os trabalhos terminológicos de áreas específicas, ou, ainda, apenas para checar dúvidas. Para LEWANDOWSKA-TOMASZCZYK, "in dictionary making they [*corpora*] are the optimal reference material against which a lexicographer's intuition can be checked up" (LEWANDOWSKA-TOMASZCZYK 1997: 254).

A fraseologia – subárea da lexicologia que se dedica ao estudo de provérbios, ditos populares, expressões idiomáticas, frases feitas, jargões, colocações etc. – também tira proveito da utilização de *corpora*. Principalmente em relação ao estudo das colocações, conforme vimos anteriormente, tal utilização se configura como um instrumento de grande utilidade. Isso porque as ferramentas de busca em *corpora* podem localizar e listar rápida e facilmente as colocações em que uma determinada palavra aparece, fornecendo dados estatísticos sobre sua freqüência e sobre a freqüência das outras palavras que a acompanham.

Esse mecanismo permite, a partir da observação e da quantificação, estabelecer que, dentro de um determinado âmbito técnico, certa colocação representa a regra geral, a norma, já que se consubstancia como a colocação mais utilizada pelos membros daquela área, ao passo que outra, ainda que também usada, representa um desvio da norma, uma variante, dado o caráter raro de sua ocorrência.

7.3. Pesquisa de padrões gramaticais e de tendências da língua

As áreas de pesquisa voltadas ao estudo de padrões gramaticais podem igualmente valer-se de corpora como um recurso proveitoso. O citado trabalho desenvolvido por BIBER (1998) é um bom exemplo de como é possível extrair do *corpus* dados informativos sobre padrões gramaticais.

Lynne Bowker (1998) explica que é muito difícil detectar alguns padrões lingüísticos quando os mesmos se encontram espalhados em um texto ou em vários textos. A elaboração de uma concordância é um meio rápido de se juntar todas as ocorrências de um dado padrão. E, acrescentamos, de quantificar sua freqüência.

No âmbito dos estudos gramaticais, o *corpus* funciona como uma ferramenta que complementa a introspecção do pesquisador com a observação empírica da língua, fornecendo-lhe subsídios para formalizar e fundamentar algumas regras que intuía.

Um exemplo de pesquisa com o alemão pode ser dado pelo levantamento que fizemos em um *corpus* a partir do vocábulo *überhaupt*. Procurando pelo *Altavista* em *sites* da *Web*, localizamos 127 ocorrências do vocábulo. Dessas ocorrências, 46 eram em frases interrogativas, o que é um número bastante significativo, já que representa quase 40% do total. Esse fato pode ser considerado como o indício de um certo padrão gramatical, o qual, porém, não analisaremos no presente trabalho.

Utilizando-se *corpora* de língua falada é possível observar a dinâmica da língua, as tendências que estão se delineando. Mediante a utilização de um *corpus* com anotações detalhadas sobre os produtores do discurso, um estudo, por exemplo, do fenômeno do gerundismo no português atual do Brasil ficaria muito enriquecido, até porque poderia delimitar exatamente o grupo de pessoas e as regiões do país em que ele mais ocorre.

Um estudo interessante sobre as tendências lingüísticas, a partir do uso de *corpus* de língua falada, foi realizado por Kjellmer (1999), pesquisando o verbo *try*, do inglês. Por meio da observação dos dados, ele detectou uma tendência desse verbo a funcionar como verbo auxiliar, já que, na linguagem falada, ele vem gradativamente perdendo seu papel lexical.

7.4. Estudos contrastivos de línguas

Inúmeros estudos contrastivos de língua podem ser desenvolvidos a partir do uso de *corpora*. Um exemplo encontrado em Johansson (1997) ilustra bem a questão. Utilizando um *corpus* multilíngüe, ele desenvolveu um estudo contrastivo do pronome genérico *one*, do inglês, com os pronomes correspondentes do alemão e do norueguês, *man*. De acordo com a tabela comparativa, montada a partir do *corpus*, constatou que o pronome *one*, do inglês, é muito menos freqüente do que o *man*, do alemão e do norueguês. Isso porque o inglês emprega outros pronomes pessoais genéricos para se referir a pessoas em geral, especialmente o *you*.

Outro exemplo que mostra como o trabalho com *corpora* fornece subsídios para as pesquisas contrastivas, é o seguinte: utilizando um *corpus* bilíngüe português-alemão, foi realizado um levantamento de ocorrências do advérbio *talvez*, em português, para verificar sua ocorrência junto com o modo subjuntivo. Os resultados mostraram claramente o emprego do modo subjuntivo nas frases em que aparece o advérbio *talvez*, expressando a idéia de probabilidade. Fazendo buscas com os advérbios *provavelmente* e *possivelmente*, muitas vezes usados como sinônimos de *talvez*, não foi encontrado, entretanto, o uso do subjuntivo. Com o intuito de comparar este tipo de uso do modo subjuntivo em português com o alemão, foram feitas buscas pelas palavras *vielleicht, wahrscheinlich* e *möglicherweise* nos textos em alemão. A análise das concordâncias encontradas mostra que estes advérbios não exigem o uso do modo subjuntivo em alemão (Gatti, 2001).

7.5. Tradução

O uso de *corpora* para a tradução é igualmente de grande valia, pois permite o acesso fácil e rápido a textos de especialidade, nos quais é possível pesquisar os termos e as expressões específicas, correntes dentro das áreas técnicas.

Para o tradutor, um dos requisitos mais importantes para a elaboração de uma boa tradução é conhecer o assunto do texto a ser traduzido. Dessa forma, o auxílio de uma coleção de textos sobre tal assunto, permitir-lhe-á uma noção geral e rápida do tema e irá familiarizá-lo com a terminologia daquele âmbito.

Segundo um estudo piloto desenvolvido por Bowker (1998), com estudantes de tradução, em que foi utilizada a pesquisa com *corpora*, o *corpus* "tem o potencial de ajudar os estudantes a encontrar e a utilizar os termos corretos." (Bowker 1998: 641)

Assim, a obtenção da acuidade na escolha dos termos na tradução de textos de uma determinada área de especialidade pode ser fornecida pela pesquisa em *corpus*. Tal acuidade contribui para o aprimoramento da linguagem empregada no texto traduzido, conferindo-lhe caráter natural.

Para ilustrar, mencionamos uma experiência que tivemos com a tradução do termo "produto final" para o alemão. Em nenhum dicionário, mesmo nos de economia, foi encontrado o registro desse termo. A dúvida era se o vocábulo em alemão seria *Schlussprodukt* ou *Endprodukt*. Em um *corpus* da *Internet*, procuramos pela primeira opção, dela porém não encontrando uma ocorrência sequer. Já o termo *Endprodukt* registrou 102 ocorrências e, pelo contexto, pudemos detectar que sua acepção era equivalente à do termo em português.

7.6. Estudos de literatura

Embora possa parecer pouco usual, até mesmo os estudos de literatura podem se beneficiar com as vantagens da pesquisa em *corpus*.

Vejamos um caso concreto, mencionado por KETTEMANN (1997), sobre um estudo realizado a partir de um conto americano de Mary Freeman Wilkins, denominado "The Revolt of Mother". No conto, a caracterização dos papéis sociais é apreensível mediante a observação dos verbos associados à personagem feminina. Eles estão sempre relacionados ao tradicional papel da dona de casa, como *cozinhar, limpar, costurar, lavar* etc. Já os verbos associados ao personagem masculino são de outro grupo semântico e designam o indivíduo que controla a situação. São eles, *designar, estabelecer, planejar, pensar*, dentre outros. No decorrer da história, há uma mudança no comportamento da personagem feminina, com a conseqüente alteração do grupo semântico dos verbos ligados a ela.

7.7. Prosódia semântica

Segundo HOEY (1997), certos usos de palavras e frases demonstram uma tendência a ocorrerem em determinados ambientes semânticos. Por exemplo, o verbo *happen*, do inglês, está associado a fatos desagradáveis e infelizes – acidentes e coisas do gênero. Assim, o uso de *happen* prepara o ouvinte/leitor para a recepção de algo ruim ou infeliz.

Esse fenômeno ocorre em todas as línguas. Por exemplo, no alemão temos o verbo *verüben*, que apresenta uma prosódia semântica negativa, já que em todas as suas ocorrências está associado a fatos ruins como crime, delito, erro etc. (assim como *begehen*). Já o substantivo *Lebensabend* tem uma prosódia semântica positiva, encontrando-se normalmente associado a outros vocábulos de ambientes semânticos positivos, como *geniessen, schön* etc., conforme averiguamos em *corpora* do *Institut für Deutsche Sprache*. Registre-se, porém, que tanto o dicionário *Langenscheidt* quanto o dicionário do Porto fazem constar como equivalente a *Lebensabend* o vocábulo português *velhice*, o que consideramos uma impropriedade, já que a prosódia semântica do vocábulo *velhice* é negativa e não resgata de forma alguma o caráter suave e positivo do vocábulo alemão.

O verbo *contrair*, do português, tem – por sua vez – uma forte associação a fatos negativos, visto ser mais freqüentemente acompanhado de palavras ligadas a eventos negativos, como doenças, dívidas, obrigações etc. Ainda assim, em contextos lingüísticos específicos, como da área jurídica, pudemos encontrar o registro da expressão *contrair matrimônio* (SOUZA, 2001).

Em outro exemplo, SARDINHA (2000), detectou que a expressão "tocar para frente", do português do Brasil, está fortemente associada à superveniência de adversidades[2].

Existem também vocábulos de prosódia semântica neutra. Eles não evocam, de antemão, nenhuma idéia ou sensação específicas, ficando a cargo do contexto a confirmação e o direcionamento de sua prosódia semântica.

8. Considerações finais

Conforme pudemos observar a partir de todo o anteriormente exposto, a Lingüística de *Corpus* oferece inúmeros recursos à pesquisa em língua, lingüística, tradução e, até mesmo, literatura. Ela dispõe de ferramentas que facilitam a elaboração estatística, a quantificação de dados, a observação de co-ocorrências. Além disso, proporciona o acesso ao contexto integral do qual cada fragmento foi extraído, fornece listas de palavras, indica quais são as de maior freqüência etc. Enfim, trata-se de instrumento valioso para o pesquisador de língua.

---

[2] Este exemplo foi extraído de uma palestra apresentada por Tony Berber Sardinha na disciplina de Pós-Graduação sobre Lingüística do *Corpus*, ministrada pela Profª. Drª Stella Tagnin, na USP, no segundo semestre de 2000.

Referências bibliográficas

AARTS, Jan. "Intuition-based and observation-based grammars". In: AIJMER, K. & ALTENBERG, B. (eds.). *English Corpus Linguistics*. London/New York, Longman 1991, 44-61.

BIBER, Douglas, CONRAD, Susan & REPPEN, Randi . *Corpus Linguistics – Investigating language structure and use*. Cambridge, Cambridge University Press 1998.

BOWKER, Lynn. "Using specialized monolingual native-language corpora as a translation resource: a pilot study". In: Meta XLIII, 4, 1998, 631-651.

GATTI, Iris Kurz. "A Lingüística do *Corpus* e os estudos de língua ilustrados por meio de exemplos do modo subjuntivo em alemão e em português". EPLLE 2001 (no prelo).

GECKELER, H. *Semántica estructural y teoria del campo léxico*. Madrid, Gredos 1984.

GENOUVRIER, Emile, PEYTARD, Jean. *Lingüística e ensino do português*. Coimbra, Almedina 1974.

GREIMAS, A. J., COURTÈS, Joseph. *Dicionário de Semiótica*. São Paulo, Ed. Cultrix 1979.

HOEYI, Michael. "From concordance to text structure: new uses for computer *corpora*". In: Lewandowska-Tomaszczyk, B. & P. J. MELIA (eds.). *PALC'97 Practical Applications in Language Corpora*. Lodz, Lodz University Press 1997, 2-23.

JOHANSSON, Stig. "Using the English-Norwegian parallel *corpus* – a *corpus* for contrastive analysis and translation studies". In: Lewandowska-Tomaszczyk, B. & P. J. Melia (eds.). *PALC'97 Practical Applications in Language Corpora*, Lodz, Lodz University Press 1997, 282-296.

KETTEMANN, Bernhard. "Concordancing as input enhancement in ELT". In: Lewandowska-Tomaszczyk, B. & P. J. Melia (eds.). *PALC'97 Practical Applications in Language Corpora*. Lodz, Lodz University Press 1997, 63-73.

KJELLMER, Göran. "Auxiliary Marginalities: The Case of 'Try'". Separata. Amsterdam – Atlanta, Rodopi 1999.

LEWANDOWSKA-TOMASZCZYK, B. & P. J. Melia (eds.). "Lexical meanings in language corpora". *PALC'97 Practical Applications in Language Corpora*. Lodz, Lodz University Press 1997, 236-255.

SALKIE, Raphael. "Naturalness and contrastive linguistics". In: Lewandowska-Tomaszczyk, B. & P. J. Melia (eds.). *PALC'97 Practical Applications in Language Corpora*. Lodz, Lodz University Press 1997, 297-312.

SANCHEZ, A. Definición e historia de los *corpus*. In: SANCHEZ, A. et al. (org.). *CUMBRE – Corpus Lingüístico de Español Contemporáneo*. Madrid, SGEL 1995.

SARDINHA, T. B. "O que é um *corpus* representativo?". São Paulo, PUC 1999 (manuscrito não publicado).

SOUZA, Eurides Avance de. "A Lingüística do *Corpus* e os estudos de língua alemã nas áreas de fraseologia, lexicografia e tradução". EPLLE 2001 (no prelo).

TAGNIN, Stella E. O. "COMET – Um *Corpus* Multilíngüe para Ensino e Tradução". São Paulo, Universidade de São Paulo 2001 (manuscrito não publicado).

Dicionários

Dicionário de Alemão Português. Porto: Porto Editora 1997.

IRMEN, Friedrich & KOLLERT, Ana Maria Cortes. *Langenscheidts Taschenwörterbuch der Portugiesischen und Deutschen Sprache*. Berlin/München/Wien/Zürich, Langenscheidt 1982.

Sites importantes

IDS – Institut für Deutsche Sprache
http://corpora.ids.mannheim.de

Corpus do IME-USP
www.ime.usp.br/~tycho

Anotador de textos
http://visl.hum.sdv.dk

Ferramentas de elaboração de concordâncias: Wordsmith Tools
www.liv.ac.uk/~ms2928/
http://www.webcorp.org.uk/