Novos horizontes para o ensino do léxico

Anna Maria Becker Maciel*

Resumo: Meu objetivo nesse artigo é chamar a atenção do professor de línguas para a importância do Corpus Lingüístico como um campo de estudos que questiona antigas crenças relacionadas à língua especificamente no que concerne ao ensino tradicional do vocabulário e da gramática. Primeiramente, fornecerei algum background geral sobre a área. Então, caracterizarei seu propósito tão bem quanto suas bases teóricas e metodológicas que enfatizam a relação entre os significados das palavras e os padrões sintáticos e coloquialismos onde estes ocorrem. A seguir, destacarei a diferença entre dois campos de pesquisa lingüística que dependem de recursos computacionais, o Corpus Lingüístico e a Lingüística Computacional. Finalmente, comentarei sobre a contribuição potencial que o enfoque do Corpus Lingüístico pode trazer para o ensino de línguas.

Palavras-chave: Corpus Lingüístico, lexicogramática, ensino de línguas, corpus, concordância.

Abstract: My aim in this paper is to draw the attention of the language teacher to the importance of Corpus Linguistics as a field of linguistic studies that calls into question long-held beliefs about language specially the traditional breakdown into vocabulary and grammar upon which much of language teaching is based. First, I will provide some general background about the area. Then, I will characterize its purpose as well as its theoretical and methodological bases which stress the relationship between the meanings of words and the syntactical patterns and collocations in which they typically occur. Next, I will outline the difference between two fields of linguistic research that depend on computational resources, Corpus Linguistics and Computational Linguistcs. Finally,

* Grupo TERMISUL-PPG LETRAS, UFRGS. ambmaciel@terra.com.br

I will comment on the potential contribution which Corpus Linguistics approach can bring to language teaching
Key-words: Corpus Linguistics, lexicogrammar, language teaching, corpus, concordance.

1 Introdução

A introdução da informática na escola desde as séries iniciais tem trazido vantagens indiscutíveis para o ensino da língua. Cada vez mais surgem novas idéias, desenvolvem-se pesquisas e discutem-se propostas. No entanto, muitas vezes, o que parece um avanço, nada mais é do que a mesma metodologia do lápis, papel, quadro negro e giz transportada para o meio digital. Talvez tal aconteça porque, levados pelo afã da novidade tecnológica, os professores desconheçam a Lingüística de *Corpus* e sua contribuição para a competência e o desempenho lingüístico do aprendiz. Por isso, tendo em mente o rico potencial pedagógico-didático dessa contribuição, teço aqui algumas considerações.

Meu objetivo é despertar a curiosidade do professor para a imensa gama de possibilidades que podem ser exploradas para orientar o aluno no processo de descoberta das riquezas da língua. Trata-se apenas de um breve panorama introdutório a uma área de estudos lingüísticos ainda ignorada nos cursos de graduação em Letras e pouco desenvolvida nos programas de pós-graduação no Brasil.

Em primeiro lugar, busco esclarecer o que é a Lingüística de C*orpus*, uma vez que, embora já conhecida desde a segunda metade do século XX, somente nos últimos anos, ela passou a ser valorizada no nosso país. Em seguida, apresento suas bases teórico-metodológicas e ferramentas principais. Depois, sintetizo as diferenças entre os dois campos de estudos lingüísticos que recorrem à informática, Lingüística de Corpus e Lingüística Computacional. Finalmente, enfatizo a contribuição da Lingüística de *Corpus* para o ensino do léxico, mesmo para aqueles professores de escolas que não dispõem de grandes verbas e equipamentos sofisticados.

2 O que é a lingüística de corpus?

A Lingüística de Corpus parece assustar aqueles que se dedicam às Letras, o termo é imediatamente associado à Informática, a fórmulas matemáticas, a conceitos estatísticos e a teoria da probabilidade. Tal é uma imagem distorcida que esconde a verdadeira face de uma lingüística, isto é, um estudo científico da língua, que utiliza um *corpus*.

Nesse momento, devo, antes de tudo, explicar o que é *corpus*. *Corpus* é um conjunto de textos digitalizados, autênticos, produzidos com o objetivo de comunicação, armazenados e preparados para pesquisa lingüística. Mas

atenção, não se pense que simplesmente trabalhar textos no computador, familiarizar-se com a Internet, transitar pela a grande rede mundial de informações, a chamada "web", seja Lingüística de *Corpus*. Lingüística de *Corpus* é a área de estudos lingüísticos que analisa os padrões de uso real da língua em grandes conjuntos de textos reais, observando empiricamente quais as formas gramaticais possíveis e prováveis de serem ditas pelos falantes de carne e osso e não por potenciais falantes idealizados.

Vale lembrar que, antes do advento dos computadores, a língua era pesquisada em pequenas amostragens artificialmente produzidas ou em exemplos autênticos de poucas de frases recolhidas especificamente com o propósito de comprovar o uso prescrito por regras gramaticais. Hoje, a Lingüística de Corpus possibilita que a língua seja analisada em contextos de centenas de milhões de textos produzidos naturalmente na interlocução dos falantes em pleno evento comunicativo sem nenhuma intenção de ilustrar um fato lingüístico.

Os exemplos de uso que os teóricos elaboravam ou recolhiam de informantes especialmente convocados e tecnicamente interrogados podem ser agora comparados com amostragens aleatórias colhidas em milhares de realizações textuais. Os autores de dicionários foram os primeiros beneficiados pelo advento da Lingüística de *Corpus*. Atualmente a edição de dicionários já não depende do olho nu dos seus autores, pois é auxiliada pelo computador no exame de extensas coleções de textos.

Para citar exemplos nossos, menciono duas publicações recentes: o Dicionário de Usos do Português do Brasil[1] e o Dicionário UNESP do Português Contemporâneo[2] elaborados a partir das entradas mais recorrentes de um universo de 90 milhões de itens lexicais extraídos do banco de dados do Laboratório de Lexicografia da Faculdade de Ciências e Letras de Araraquara que reúne textos escritos no Brasil de prosa romanesca, dramática, técnica, oratória e jornalística com absoluta predominância desta última de 1950 até hoje.

3 Bases teórico-metodológicas

As bases teórico-metodológicas da Lingüística de Corpus podem ser encontradas nos trabalhos do britânico J.R. Firth (1980-1960) que, em um computador gigantesco, verdadeiro dinossauro comparado aos nossos *laptops*, nos anos 50 já pesquisa em textos autênticos a distribuição de palavras sócio-culturalmente

[1] BORBA, F. *Dicionário de usos do Português do Brasil*. São Paulo: Ática, 2002. 1674p.

[2] BORBA, F. *Dicionário UNESP do Português Contemporâneo*. São Paulo, UNESP, 2005. 1474p.

relevantes, porque acredita que o significado se configura no contexto de uso. Sua tão repetida citação "*You shall know a word by the company it keeps*" chama atenção para a imensa rede de relações sintagmáticas e paradigmáticas que envolve léxico e gramática, apontando para o fenômeno que ele chama colocação. Observa, também que as palavras que o falante escolhe do potencial lingüístico a sua disposição têm um padrão de associação regular. isto quer dizer que as palavras privilegiam um tipo de combinação ou, melhor dito, elas preferem determinados vizinhos ou ainda rejeitam certas estruturas.

Para ilustrar tal fenômeno, menciono uma pesquisa de Berber Sardinha (1999) que observa, em um corpus de mais de 32 milhões de palavras, as quatro palavras colocadas antes e depois do verbo "causar". O pesquisador verifica que a grande maioria das palavras colocadas ao redor desse verbo, diz respeito a eventos ou aspectos negativos, tais como problemas, danos, morte, prejuízos, doenças, males.

A regularidade expressa na recorrência sistemática de unidades co-ocorrentes de várias ordens, léxica, sintática, gramatical e semântica, é objeto de estudo da Lingüística de *Corpus* que distingue três tipos de padrões léxico-gramaticais: colocação, coligação e prosódia semântica. Colocação é a associação entre itens lexicais Coligação é associação entre itens lexicais e gramaticais. Prosódia semântica é associação entre itens lexicais e conotação (negativa, positiva ou neutra) de campos semânticos.

Intuição e introspecção não se contrapõem à observação empírica defendida pela Lingüística de Corpus. A famosa caricatura do lingüista de gabinete e do lingüista de *corpus* delineada por Fillmore (1992: 35-59) bem caracteriza o inter-relacionamento necessário entre as duas posições. Enquanto o lingüista de gabinete fica em sua poltrona meditando, mergulhado na sua intuição de falante nativo, o lingüista de corpus fica diante do computador mergulhado em milhares de palavras, imerso em milhões de textos. No entanto, continua Fillmore, não são dois pesquisadores em contradição, são dois lingüistas que estudam o mesmo objeto e que devem co-existir em uma mesma pessoa, pois competência e desempenho não se excluem. Evidência e intuição são indispensáveis na análise lingüística, configurando duas faces de uma mesma moeda. Leech (1995:74) acrescenta que não existe antagonismo entre as posições desses lingüistas, não se trata de optar por racionalismo ou por empiricismo, mas de unir a introspecção e a observação empírica da língua em uma mesma pessoa.

Os dados que o *corpus* oferece não se contrapõem aos dados da intuição do falante nativo, no entanto, muitas vezes evidenciam fatos sobre o uso real da língua que talvez nenhum falante nativo e nenhum lingüista teriam imaginado. Tais fatos, longe

de serem ocorrências incorretas de desvios de formas não gramaticais, são realizações que exemplificam padrões estruturais, distribucionais, propriedades léxicas e discursivas imperceptíveis em amostragens de extensão reduzida.

Os procedimentos matemáticos empregados na pesquisa, não reduzem a Lingüística de Corpus a um mero exercício matemático, tampouco a uma quantificação de fatos lingüísticos. O recurso aos procedimentos estatísticos se fundamenta no caráter probabilístico da língua comprovado teoricamente por lingüistas e matemáticos e, empiricamente, pelas realizações dos falantes em situações reais de comunicação. Além disso, a estatística se explica pela necessidade de avaliar qual é a significância dos dados coletados em grandes extensões de textos e assim poder objetivamente determinar seu valor como amostragem do sistema da língua.

4 As ferramentas da lingüística de corpus

A Lingüística de *Corpus* não é um conjunto ferramentas computacionais. Seu instrumental é um meio de aplicar os pressupostos teóricos que partem da descrição empírica da língua em uso evidenciada em grandes conjuntos de textos criteriosamente coletados. Sua proposta metodológica exige o uso do computador, no entanto, o processamento da pesquisa longe de ser automatizado, é interativo. O pesquisador é quem determina os dados relevantes a extrair, conduz a análise, levanta hipóteses, compara resultados, reflete, discute e chega a conclusões. Dessa maneira, conjugam-se homem e máquina: de um lado, o conhecimento científico e a intuição da língua e do outro, a informática e a tecnologia

Os *corpora* são examinados com o auxílio do computador, pois são tão extensos que ninguém jamais os poderia analisar manualmente, mesmo que existisse uma equipe gigantesca. Tal equipe poderosa seria vulnerável à fadiga e à subjetividade próprias da natureza humana. Ao passo que, a máquina, uma vez alimentada de maneira inteligente pelo homem, é sempre a mesma, incansável e consistente em seus achados. Os computadores pessoais, os PCs, tornaram acessível a todos a automatização dos processos de estocagem, recuperação, gerenciamento e leitura de textos. A *web* chega aos mais recônditos confins do universo e coloca à disposição de quem quer que seja uma infinidade de textos nos quais a língua em uso na efetiva comunicação humana pode ser analisada sob mais variados pontos de vista.

A principal ferramenta da Lingüística de Corpus é o *software* capaz de fazer concordâncias, o Concordanciador. É preciso esclarecer aqui o conceito de concordância que não tem nada a ver com o conceito gramatical

que conhecemos. A concordância é uma listagem de contextos em que um dado item, que pode ser uma palavra isolada, uma palavra composta, uma estrutura, um sinal de pontuação, aparece no centro da linha, ladeado pelas palavras que ocorrem antes e depois dele no texto. Tal procedimento característico da Lingüística de *Corpus* não é nada novo, os monges do século XII já o utilizavam manualmente para listar as palavras da Bíblia. Hoje a tecnologia da informática tornou possível sua realização com um simples clicar de teclado. A novidade da concordância moderna é a facilidade de sua operacionalização oportunizada pelo computador e que é aproveitada para abrir caminhos antes insondáveis para a pesquisa lingüística.

As principais ferramentas utilizadas na lingüística de corpus se encontram à disposição do usuário na *web*, algumas livres de qualquer taxa, outras comercialmente distribuídas. Da mesma maneira, os grandes *corpora* referenciais aí estão e podem ser acessados *on-line* na maioria dos idiomas. A disponibilização de recursos e a divulgação das pesquisas derivam do aspecto social da comunicação um dos princípios básicos que lingüística de *corpus* enfatiza.

5 Lingüística de *corpus* e lingüística computacional [3]

Lingüística de *Corpus* e Lingüística Computacional são duas áreas distintas que não se opõem, mas que têm campos de estudo diversos, princípios e objetivos diferentes. A primeira está voltada exclusivamente para a descrição da língua como evento social comunicativo e encontra em Firth, conforme anteriormente mencionado, e na teoria sistêmica funcional desenvolvida por Halliday, apoio para sua metodologia. Já a Lingüística Computacional se ocupa da descrição da língua em função da construção de sistemas automatizados com capacidade de reconhecer e produzir a linguagem natural. Para tanto, a Lingüística Computacional vê a língua como um conjunto de regras finitas e aproxima-se das teorias lingüísticas racionalistas que dão primazia á competência do falante e concede prioridade ao processamento da língua natural (PLN).

Assim, enquanto a Lingüística de *Corpus* enfoca a comunicação dos falantes humanos entre si, a lingüística computacional, como área multidisciplinar, visa de maneira especial à comunicação do homem com a máquina, estuda os processamentos sintático,

[4] Para maiores esclarecimentos ver as entrevistas de Renata Vieira e de Berber Sardinha em *Revista Virtual de Estudos da Linguagem - ReVEL*. Ano 2, n. 3. [www.revelhp.cjb.net]

semântico e lógico da linguagem natural. Sua motivação pode ser simplesmente teórica voltada para a ciência cognitiva para explicar fenômenos psicolingüísticos através de regras e algorítmos, ou pode ser prática, na perspectiva da engenharia lingüística, de modo a contribuir para a construção de sistemas automatizados, tais como inteligência artificial (IA), reconhecimento da voz, tradução automática, geração de resumos, recuperação da informação, entre outros.

6 Lingüística de *corpus* e ensino do léxico

A Lingüística de *Corpus* abre novos horizontes para os procedimentos de ensino-aprendizagem da língua, tanto materna, quanto estrangeira. Nessa área, a influência estruturalista se faz sentir até hoje e determina freqüentemente a separação do ensino do vocabulário do ensino da gramática. No entanto, as pesquisas realizadas à luz da Lingüística de *Corpus* não permitem tal dicotomia, uma vez que a descrição da língua fundamentada na observação empírica da realidade dos *corpora* evidencia que a palavra e as regras de uso não estão em dois níveis diferentes. Ao contrário, léxico e gramática integram um único nível do sistema lingüístico e sua abordagem não pode ser dissociada nem na comunicação real, nem nos ensinamentos da sala de aula.

Nesse contexto, a Lingüística de *Corpus* abre novos caminhos para que professor e aluno percebam, a partir de realizações textuais autênticas, a complexidade do inter-relacionamento do léxico, da sintaxe e da semântica e possam fazer suas descobertas, selecionando elementos lexicais e regras gramaticais de acordo com significado que desejam expressar na comunicação. Em tal integração, torna-se possível desenvolver a conscientização lingüística e a autonomia do aluno no uso da língua, tão valorizadas no processo pedagógico-didatico da comunicação lingüística.

Para que esse horizonte seja alcançado, é preciso que, em primeiro lugar, que o professor conheça os pressupostos básicos da proposta da lingüística de corpus e procure se familiarizar com a metodologia e as ferramentas adequadas para sua implementação. Tal introdução está ao alcance de todos nas páginas *web* que oferecem desde ensinamentos para iniciantes até sugestões para pesquisadores avançados. A socialização dos recursos e dos achados da lingüística de *corpus* é um dos pilares que a sustentam e comprovam o pressuposto de que a língua é um comportamento social antes que propriedade privada e uma habilidade individual.

Como toda a perspectiva nova, a Lingüística de *Corpus* suscitou e ainda suscita muitos debates, a favor e contra. Muitos lingüistas de grande projeção, tais como Chomsky, simplesmente a rejeitam, outros a vêm como uma disciplina, enquanto outros a consideram apenas

uma metodologia. Sem entrar na discussão, recorro a Berber Sardinha (2004, 36-38) que registra a opinião de destacados lingüistas que preferem dizer que a Lingüística de C*orpus* é uma abordagem para os estudos lingüísticos antes que um ramo da lingüística. Nesse direcionamento, a Lingüística de *Corpus* abre um leque infinito de possibilidades para a renovação dos estudos lingüísticos e a construção da Lingüística do nosso século XXI.

Referências Bibliográficas

BERBER SARDINHA, T. **Lingüística de Corpus**. São Paulo: Manole, 2004

BERBER SARDINHA, T. Computador, corpus e concordância no ensino de léxicogramática de língua estrangeira. In V. Leffa (Ed.), **As Palavras e sua Companhia – O Léxico na Aprendizagem**, p.42-72.. Pelotas, RS: EDUCAT / ALAB.

FILLMORE, C.J. "Corpus linguistics" or "Computer-aided armchair linguistics". In: SVARTVIK, J. (ed.) **Directions in Corpus Linguistics**. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin: Mouton de Gruyter, 1992. (Trends in Linguistics/Studies and Monographs). p. 35-59.

LEECH, G. Corpora. In: MALMKJAER, K. (ed.) **The Linguistics Encyclopaedia**. London: Routledge, 1995. p. 73-80