Apresentação

Linguística de *Corpus* no Brasil: uma aventura mais do que adequada

Giovanni Parodi (2010, p. 167), nas reflexões finais de seu livro *Linguística de Corpus: da teoria à empiria*, trazia, quatro anos atrás, esta impressão: "parecem tempos em que ser linguista de *corpus* é uma aventura adequada". Considerando o percurso da Linguística de *Corpus* (doravante LC) no nosso país desde 2004, época do lançamento do livro de Tony Berber Sardinha, nosso primeiro manual brasileiro de LC, cabe aqui refletir um pouco sobre essa combinação de palavras dirigida à Linguística de *Corpus*.

Aventuras podem ser mais ou menos adequadas? O que é ser adequado em termos de uma aventura no terreno dos Estudos da Linguagem no nosso país? Ao apresentar este volume da revista Letras & Letras, queremos também tratar dessas questões.

Neste ano de 2014, comemoramos também os cinquenta anos do *Corpus Brown* (1964), um ponto de referência inevitável em qualquer retrospectiva sobre a LC em nível mundial. Ainda que de modo bastante mais restrito, em termos de repercussão local, podemos destacar que, exatamente há um mês, em novembro de 2014, conseguimos completar mais uma aventura, realizamos o *XII Encontro de Linguística de Corpus (ELC)* e a *VII Escola Brasileira de Linguística Computacional (EBRALC)*, na Universidade Federal de Uberlândia (UFU), interior de Minas Gerais, que também recebe a organização deste número da Letras & Letras especialmente dedicado à LC.

A partir dessas trajetórias e de outras, desenhadas por toda uma comunidade de pesquisadores em LC do Brasil, as palavras do professor Parodi renovam-se em significância e não poderiam definir melhor o momento atual. Se, por um lado, essas efemérides trazem à memória alguns marcos históricos importantes e nos fazem pensar em nossos próprios percursos até aqui, conduzem também a uma reconstrução do próprio processo, na constituição da área que hoje reconhecemos como LC.

Ainda cabe repetir que a LC se coloca como uma nova perspectiva para a Linguística (BERBER SARDINHA, 2004, p. 35), mas não como um novo tipo de Linguística. Mostra-se, para aqueles que se aproximam da LC, tanto como uma metodologia quanto como uma abordagem teórica diferenciada dos Estudos da Linguagem. De quem queira se aproximar da LC, apenas por se interessar por seu instrumental ou por seus procedimentos, nada será

cobrado em termos de uma filiação teórica – ou epistemológica – ainda que insistamos que LC também é um modo de compreender a língua, que temos nosso modo de defini-la como objeto de estudo: a língua é um sistema probabilístico de combinatórias, no qual uma unidade se define pelas associações que mantém com outras unidades.

Ao ocupar-se da exploração de grandes extensões de *corpora* textuais em formato digital, criteriosamente reunidos para representar um dado estado de uso de língua e "minerados" com apoio informatizado, com destaque para as explorações estatísticas de elementos lexicais e observação das frequências de combinatórias de palavras, vemos toda uma trajetória de estudos realizados no Brasil. Esses estudos, considerados em uma perspectiva muito ampla, podem ser bastante aproveitados em diferentes tipos de pesquisas e servem hoje, no mínimo, para caracterização de gêneros textuais[1]. Ao longo do seu percurso investigativo entre nós, quase todos os gêneros textuais escritos foram objeto de algum estudo em LC, do literário ao jornalístico, manuais técnicos e textos de culinária, entre vários outros, sem esquecermos dos *corpora* especialmente dedicados aos registros orais. A esse respeito, Mello (2012, p. 34) destaca que

> Apesar de os *corpora* escritos ainda dominarem a produção na área, a compilação de *corpora* orais e multimodais tem se ampliado rapidamente. Os *corpora* orais têm encontrado crescente aplicabilidade não apenas nos estudos canônicos da Linguística (Sociolinguística, Dialetologia, Lexicografia, Morfossintaxe etc.), mas também no desenvolvimento de tecnologias da fala, tais como o reconhecimento e síntese da fala.

O questionamento sobre a importância da coleta de dados dos usos linguísticos para as pesquisas, recorrente nos inícios da década dos sessenta, em pleno contexto histórico de dominância de uma linguística gerativista, contrasta radicalmente com o panorama atual. Se considerada a perspectiva de uma época em que se presumia que os dados já estariam na mente do linguista, o surgimento do *Corpus Brown* exatamente nesse contexto teve um valor pioneiro incalculável e um efeito dinamizador dos estudos baseados em *corpora*, já apontados por diversos autores. Essa mudança de paradigma se traduz num caminho percorrido entre a idealização e a sistematização da observação de evidências.

Atualmente, a expansão do uso dos termos *corpus* e *corpora,* além da menção a muitas das ferramentas e princípios caros à LC, alcança áreas que poderiam parecer, num primeiro

[1] As concepções de gênero textual presentes em trabalhos sob a ótica de Linguística de *Corpus* são em geral derivações das ideias de SWALES (1990) e HALLIDAY (1991).

momento, incompatíveis ou inimagináveis. Assim, a alusão às terminologias típicas de LC (como *types*, *tokens* e concordâncias) vem se tornando cada vez mais recorrente. Em eventos científicos, em publicações, em nomes de disciplinas, teses e dissertações, a recorrência com que aparecem referências ou vestígios da LC denotam já uma presença marcada no plano acadêmico e servem como um bom termômetro do estado da arte.

Já é amplamente conhecida a afirmação de que todo *corpus* sempre traz questões novas ou questões que não se imaginava encontrar, ainda que – de acordo com o próprio Fillmore (1992) – nenhum *corpus* nos dê resposta para tudo. De tal modo, tanto as observações como os experimentos e hipóteses formuladas no âmbito de toda investigação nos conduzem a uma revisão à luz das comprovações e dos resultados. Um médico espanhol, Ramón y Cajal, já assinalava em 1899 a importância do exame direto dos fatos da natureza e o uso de métodos, na tentativa de reduzir o máximo possível fatores subjetivos. Com isso, toda observação de dados para sua posterior descrição demandaria, necessariamente, fundamentações teóricas e princípios metodológicos; mas, acima de tudo, exigiria o traçado de caminhos de ida e de volta para a própria revisitação dos dados e ajustes dos pressupostos iniciais.

Dessa maneira, a sistematização de dados e de observações chega a ser crucial, talvez ainda mais importante do que a simples aplicação e contraste de teorias. A descoberta e identificação de padrões a partir da observação são, para Hanson (1958), os problemas fundamentais. Assim, toda teoria deriva do resultado de um trabalho consciente sobre os dados, uma vez que a tarefa das teorias seria colocar fenômenos em sistemas. Todas essas observações conduzem nosso olhar para a compreensão da relevância dada aos processos de observação, etapa indispensável nas pesquisas com *corpora* e nas diferentes fases de descoberta.

As concepções da LC, conforme vemos, com base nas ideias de Stubbs (1996, p. 46; 2001) e de Sinclair (1991, p. xviii), são as seguintes:

a) Um *corpus* não é mera ferramenta de análise. É, sim, um importante conceito teórico;

b) A linguagem se mostra diferente quando examinada extensivamente.

Assim, a "aventura adequada" citada no início deste texto envolve toda uma trajetória e todo um empreendimento coletivo de uma comunidade de pesquisadores. Nela se colocam um conceito teórico diferenciado e um empreendimento que convida o linguista interessado a

apreciar o seu objeto de estudo sob um ângulo também diferenciado. Aqui, frisamos, *diferenciado* não deve ser compreendido como algo melhor ou contrário aos diferentes convites à Linguística da nossa atualidade.

Esta apresentação busca contextualizar o momento particular em que surge este número temático da revista Letras & Letras, já definido como um marco histórico em que se constitui uma nova área de pesquisa, com abordagens e métodos próprios. Por outro lado, este texto introdutório também procura enxergar o *corpus* como essa espécie de "caminho de ida e volta", amarrado à importância da observação empírica dos fatos linguísticos. Nessa perspectiva, os *corpora* se tornam um território vasto o suficiente e propício para a descoberta de evidências.

Este número da revista Letras & Letras, dedicado à LC, está composto por 19 artigos e uma entrevista com o linguista chileno Giovanni Parodi, realizada no âmbito do XII ELC e da VII EBRALC. A diversidade de assuntos que compõe os artigos aqui presentes vai desde aspectos metodológicos a estudos de caso muito específicos, passando por diferentes correntes e afiliações teóricas, dentre as quais destacamos a Linguística Sistêmico-Funcional, a Linguística Histórica, a Documentação e a Linguística Cognitiva.

Giacomo Figueredo, da Universidade Federal de Ouro Preto (UFOP), no artigo intitulado "Uma metodologia de perfilação gramatical sistêmica baseada em *corpus*", apresenta uma metodologia de investigação de funções gramaticais, embasado em princípios da LC. O autor propõe uma metodologia que possibilita a identificação e descrição de padrões, na análise do modo como a gramática é empregada na organização do texto. Com base na teoria sistêmico-funcional e utilizando um *corpus* composto por dez minibiografias escritas em português brasileiro, o pesquisador estabelece um mapeamento das funções gramaticais, da distância topológica entre as funções e do movimento do emprego dessas funções no espaço gramatical.

Cristiane Namiuti-Temponi e Aline Silva Costa, da Universidade Estadual do Sudoeste da Bahia (UESB), também abordam aspectos metodológicos em seu texto "Reflexões sobre anotação sintática e ferramentas de busca - Uso da linguagem XML para anotação sintática no *corpus* digital DOViC". No artigo é discutido o uso da linguagem XML como alternativa ao formato *Penn TreeBank* para anotação sintática no *corpus* digital denominado *Documentos Oitocentistas de Vitória da Conquista* (DOViC). Dentre as justificativas apresentadas, as autoras destacam a utilização da linguagem XML na anotação

de edições e de informações morfológicas do *corpus*, por favorecer a criação de recursos padronizados reutilizáveis que facilitam a extração de dados dos *corpora*.

Da Universidade de São Paulo (USP), com o texto "Linguística de *Corpus* e ensino: a compilação de um *corpus* de especialidade para preparação e implementação de um curso preparatório rápido para exame de proficiência", a professora Stella Esther Ortweiler Tagnin e Danilo Suzuki Murakami apresentam o processo de compilação de um *corpus* de especialidade na área de Relações Exteriores. Os autores buscam definir tanto o conteúdo programático como a preparação de material didático para candidatos a um exame de proficiência em inglês, pensando em interessados no preenchimento de um cargo público no âmbito do governo federal.

Também no âmbito do ensino, mas voltado para a escrita em língua espanhola de aprendizes brasileiros, Benivaldo José de Araújo Júnior, da Escola Superior de Propaganda e Marketing (ESPM), apresenta o artigo "As construções com SE na produção escrita de brasileiros aprendizes de espanhol como língua estrangeira: um estudo baseado em *corpus*". Com base na Gramática Cognitiva, o autor trata especificamente das construções reflexivas, médias, impessoais e passivas e estabelece uma comparação com dados observados em dois *corpora* de falantes nativos: um de espanhol na variedade peninsular e outro de português brasileiro.

Ainda no limiar dos pressupostos teóricos da Linguística Cognitiva, Heberth Paulo Souza, do Instituto de Ensino Superior Presidente Tancredo de Almeida Neves (IPTAN), aborda a metáfora no escopo da cognição humana, no âmbito das representações mentais, com aplicações para a descrição da articulação textual. No artigo intitulado "Metáforas e domínios narrativos numa perspectiva da Linguística de *Corpus*", o autor recorre à Teoria dos Espaços Mentais e à Teoria da Mesclagem Conceitual, para descrever o papel exercido pela metáfora na articulação textual, num *corpus* de redações de vestibulandos, observando uma forma de organização de elementos típica dos processos de narração.

Para este número da revista *Letras & Letras*, os trabalhos desenvolvidos em três artigos tomam como *corpus* de estudo a legendagem. Desse modo, Vera Lúcia Santiago Araújo e Ítalo Alves Pinto de Assis, da Universidade Estadual do Ceará (UECE), no texto "A segmentação na legendagem para surdos e ensurdecidos (LSE) de Amor Eterno Amor: uma análise baseada em *corpus*", analisam problemas na divisão linguística em diálogos de uma produção audiovisual legendada. Segundo os autores, os problemas mais recorrentes de

segmentação foram detectados na ordem de sintagmas verbal e nominal, em legendas de três linhas e com alta velocidade.

Também Ana Katarinna Pessoa do Nascimento e Stella Tagnin, da USP, recorrem ao estudo das legendas, considerando a tradição francesa na LSE. No texto "Efeitos sonoros na legendagem francesa para surdos e ensurdecidos", as autoras analisam a tradução dos efeitos sonoros de três filmes franceses, com subsídios do programa *WordSmith Tools 5.0*. Por sua vez, Amanda Verdan Dib e Paulo Pinheiro-Correa, da Universidade Federal Fluminense (UFF), analisam "A função pragmática Tópico na legendagem brasileira de um filme argentino em um estudo de *corpus* paralelo", na legendagem brasileira do filme argentino *O Segredo dos seus olhos*, com a fundamentação teórica da Gramática Discursivo-Funcional (GDF) e funções do programa para alinhamento de *corpora* paralelos *YouAlign* (Terminotix Inc.). Os autores analisaram dois tipos de construções de tópicos: as topicalizações e os deslocamentos à esquerda.

Mudando o foco para o âmbito musical, Patrícia Bertoli, da Universidade Estadual do Rio de Janeiro (UERJ), apresenta o artigo "Convergência Lexical entre Letras de música e Inglês Geral: Um estudo baseado em *Corpus*". A partir de um *corpus* de aproximadamente 1 milhão de palavras, resultante de 5.962 letras de músicas diferentes, a autora contrastou listas de palavras individuais e de trigramas do *corpus* de estudo com outras listas extraídas de dois *corpora* de referência do inglês geral. Os resultados da pesquisa apontam semelhanças entre a linguagem usada nas letras das músicas analisadas e o inglês coloquial.

Por sua vez, o trabalho de Fernanda Beatriz Caricari de Morais, da Instituição Nacional de Educação de Surdos (INES), toma como objeto de estudo artigos científicos de Linguística – um estudo baseado na Linguística Sistêmico-Funcional, com o auxílio da Linguística de *Corpus*. Nele, a autora analisa os participantes agentes dos verbos do *dizer* mais utilizados em artigos de Linguística, coletados através da plataforma de periódicos *Scielo*. A LC possibilitou-lhe o tratamento computacional e dados quantitativos e contextos de ocorrência de determinadas palavras. Cláudio Márcio do Carmo, da Universidade Federal de São João del Rei (UFSJ), também na linha da Linguística Sistêmico-Funcional, combinada à Análise Crítica do Discurso, examina os discursos sobre sustentabilidade e nos traz a *ecologia linguística* uma área de trabalho em Linguística de *Corpus* que se ocupa da análise de padrões lexicais de que um determinado item faça parte, visando descrever sentidos a que esse item se associe.

Demonstrando a afinidade metodológica da LC com estudos de Linguística Histórica, o trabalho de Simone Floripi (UFU) trata de mapear o artigo do português clássico ao português europeu moderno. Sua investigação diacrônica, por meio da quantificação, mobiliza pressupostos teóricos de vertente gerativista e o Modelo de Princípios e Parâmetros. O trabalho de Maria Mercedes Riveiro Sebold e Anne Katheryne Estebe Maggessy, ambos da Universidade Federal do Rio de Janeiro (UFRJ), explora o contexto de produtividade das perífrases de gerúndio e de particípio no português do Brasil e na variedade do espanhol do México. O aspecto verbal é o ponto para contrastar essas duas línguas ricas morfologicamente.

Silvana Maria de Jesus, da Universidade Federal de Uberlândia (UFU), percorrendo outra direção, nos traz um trabalho sobre *corpus* combinado e estudos de Tradução baseados em *corpora*. No seu artigo, a autora aborda as relações de tradução de say/dizer em textos ficcionais no par linguístico inglês-português. O *corpus* combinado apresentado é composto de três romances originais em inglês e suas traduções para o português e três romances originais em português e suas traduções para o inglês, sendo parte do CORDIALL (*Corpus* Discursivo para Análises Linguísticas e Literárias) desenvolvido pelos pesquisadores do LETRA (Laboratório Experimental de Tradução) da Faculdade de Letras da UFMG.

Na perspectiva do ensino de língua estrangeira, especificamente no que se refere a padrões de escrita de aprendizes brasileiros de língua inglesa, coloca-se o trabalho de Barbara Malveira Orfano (UFSJ), Ana Larissa Adorno (UFMG) e Adriana Tenuta (UFMG), intitulado "Epistemic modality through the use of adverbs: a corpus-based study on learners' written discourse" que trata da modalidade epistêmica concretizada pelo uso de advérbios. Esse artigo discute a forma como os falantes nativos e os aprendizes brasileiros diferem em sua produção escrita e as possíveis implicações pedagógicas dessas diferenças.

Voltando ao tema de estudos em *corpora* de Tradução, o trabalho de Adriana Silvina Pagano, Arthur de Melo Sá e Kícila Ferreguetti, todos da UFMG, aborda a equivalência tradutória de partículas modais, trazendo um dos trabalhos do grupo de pesquisa "Modelagem sistêmico-funcional da tradução e da produção textual multilíngue", do Laboratório Experimental de Tradução da UFMG. Os autores compilaram um *corpus* paralelo bilíngue português brasileiro – inglês, formado por histórias seriadas da *Turma da Mônica* e suas respectivas traduções para o inglês. O objetivo foi identificar quais partículas eram utilizadas

com mais frequência no *corpus*, como elas foram traduzidas para o inglês, e seria possível verificar um padrão para as opções tradutórias.

Pedro Henrique Lima Praxedes Filho e Cristiene Ferreira da Silva, ambos da Universidade Estadual do Ceará (UECE), por sua vez, nos trazem um estudo de caso baseado em um *corpus* de roteiros de audiodescrições francesas de filmes, abordando uma das modalidades da Tradução Audiovisual – a que diz respeito à acessibilidade sociocultural de pessoas com deficiência visual. Ao abordarem o parâmetro da neutralidade em Audiodescrição, os autores buscaram investigar, com o auxílio da Linguística de *Corpus* (LC), a presença ou ausência de interpretação por parte do tradutor/audiodescritor, segundo os fundamentos da Teoria da Avaliatividade, no escopo da Linguística Sistêmico-Funcional (LSF).

Sabrina Matuda e Stella Tagnin, ambas da USP, nos trazem um artigo sobre a terminologia do futebol em um estudo direcionado pelo *corpus*. É estudada a terminologia do futebol em inglês e português, mobilizando-se a Linguística de *Corpus*, a Terminologia Textual e a concepção da tradução técnica culturalmente condicionada.

Jorge Viana Santos e Giovane Santos Brito, ambos da Universidade Estadual do Sudoeste da Bahia (UESB), brindam-nos com um trabalho bastante diferenciado, pois tratam da fotografia técnica de documentos para formação de *corpora* digitais eletrônicos. Nele, os autores apresentam as etapas do método fotográfico que vem sendo desenvolvido e utilizado, desde 2008, no Lapelinc (Laboratório de Pesquisa em Linguística de *Corpus* – UESB). Apresenta-se o processo de transposição de documentos manuscritos históricos do tipo jurídico para formação de *corpora* linguísticos.

Enfeixando o volume, temos o relato de uma entrevista feita por Ariel Novodvorski e Ana Fritz Herrera, ambos da UFU, com o linguista de *corpus* citado no início desta apresentação: o Prof. Dr. Giovanni Parodi, diretor de Pós-Graduação em Linguística da Pontifícia Universidade Católica de Valparaíso no Chile. Nessa entrevista, em novembro de 2014, durante o *XII Encontro de Linguística de Corpus (ELC)* e a *VII Escola Brasileira de Linguística Computacional (EBRALC)*, Parodi nos apresenta a plataforma de *corpora* do espanhol *El Grial*, mas também mostra-nos suas reflexões sobre o passado e o futuro da LC em seu país e em uma perspectiva global.

Em todos esses artigos, em especial na entrevista do nosso convidado, ecoam as bases teórico-metodológicas da LC, as quais remontam aos trabalhos do britânico J. R. Firth

(escritos de 1960 a 1980). Firth, lidando com um enorme computador dos anos 50, já pesquisava em textos autênticos a distribuição de palavras sócio-culturalmente relevantes e acreditava que o significado de uma palavra se configurava no contexto de uso. Sua tão repetida citação **"You shall know a word by the company it keeps"** desde então chama a atenção para a imensa rede de relações sintagmáticas e paradigmáticas que envolve léxico e gramática, apontando para o fenômeno que ele chamava de colocação. Observava Firth, como bem temos estudado em LC, que as palavras que um falante escolhe utilizar em meio a um todo de opções à sua disposição exibem um padrão de associação regular. Isto é, as palavras privilegiam um tipo de combinação ou, melhor dito, elas "preferem" determinadas associações e, ainda, "rejeitam" outras.

Assim inspirada, ao longo de sua trajetória brasileira, para finalizar este texto, cabe uma analogia com a célebre citação de Firth e com as colocações, situando a LC no nosso cenário de estudos linguísticos. Conforme acreditamos, a LC associou-se a diferentes aventuras de investigação e praticamente nada rejeitou em termos de parcerias de trabalho – o diálogo tem sido uma marca constante, mesmo com aqueles que encaram a LC apenas como um *modus operandi* computacional e quantitativo. A despeito dessa impressão, claro deve ter ficado nesses, pelo menos, primeiros 10 anos de percurso no Brasil, que vamos muito além de "contar palavras" e que já prestamos uma contribuição muito importante para toda uma comunidade de pesquisa nacional e globalmente conectada. Assim, a aventura tem sido, sim, adequada e, mais do que isso, já muito bem-sucedida.

Desejamos a todos uma ótima leitura dos trabalhos deste volume e agradecemos ao nosso colega Prof. Dr. Guilherme Fromm, pelo suporte sempre atento a tudo que precisamos durante a organização deste número da Letras & Letras.

Ariel Novodvorski[*]
Maria José Bocorny Finatto[**]
Editores

[*] Doutor em Estudos Linguísticos pela Universidade Federal de Minas Gerais (UFMG). Professor Adjunto no Curso de Graduação em Letras e no Programa de Pós-Graduação em Estudos Linguísticos (PPGEL) do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU).
[**] Doutora em Letras (UFRGS, 2001). Docente do Departamento de Linguística, Filologia e Teoria Literária da UFRGS, orientadora de mestrado e de doutorado junto ao PPG-Letras da UFRGS na linha de pesquisa Lexicografia e Terminologia: Relações textuais, Especialidade: Teorias do Léxico. Coordenadora do GELCORP-SUL, Grupo de estudos em Linguística de Corpus do Sul, certificado pela UFRGS, pesquisadora do grupo TERMISUL, bolsista PQ-CNPq.

Referências Bibliográficas

BERBER SARDINHA, A. **Lingüística de Corpus**. Barueri, SP: Manole, 2004.

FILLMORE, C. J. Corpus linguistics or computer corpus linguistics. In: SVARTVIK, J. (org). **Directions in corpus linguistics. Proceedings of Nobel Symposium 82**, *Stockholm.* Berlim/Nova York, De Gruyter, 1992.

HALLIDAY, M. A. K. Corpus studies and probabilistic grammar. In: AIJMER, K. ; ALTENBERG, B. (Org.). **English corpus linguistics**: studies in honour of Jan Svartvik. London: Longman, 1991.

HANSON, N. **Patrones de descubrimiento**: observación y explicación. Madrid: Alianza Editorial, 1977 [1958].

PARODI, G. **Lingüística de Corpus**: de la teoría a la empiria. Madrid: Iberoamericana / Vervuert, 2010.

RAMÓN Y CAJAL, S. **Reglas y consejos sobre investigación biológica**. Segunda edição do seu discurso lido perante a R.A.C.E.F. e N. Madrid: Imprenta de Fortanet, 1899.

RASO, T.; MELLO, H. (org.). **C-ORAL-BRASIL I**: Corpus de referência do português brasileiro falado informal. Belo Horizonte: UFMG, 2012.

SÁNCHEZ, A. (org.). **Cumbre – Corpus Lingüístico del español contemporáneo**: fundamentos, metodología y aplicaciones. Madrid: SGEL, 1996.

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

STUBBS, M. **Text and corpus analysis**: Computer-assisted studies of language and culture. Oxford: Blackwell, 1996.

SWALES, J. M. **Genre analysis**: English in academic and research settings. Cambridge: Cambridge University Press, 1990.