

Apresentação

Linguística de *Corpus*, metodologias e interfaces

É com prazer que apresentamos ao público leitor da Revista Domínios de Lingu@gem o seu primeiro número temático de 2015, voltado para a pesquisa em Linguística de *Corpus*, ou aquela que utiliza a metodologia proposta por este campo disciplinar da Linguística como ferramenta de investigação. Conforme apontado por Fromm e Novodvorsky (2015), a Linguística de *Corpus* vem ganhando espaço nos periódicos científicos brasileiros nos últimos anos, notadamente nas seguintes publicações temáticas, inteiramente dedicadas aos relatos de avanços na pesquisa da área: Veredas 13:2 (2009), Revista Brasileira de Linguística Aplicada 11:2 (2011) e Letras e Letras 30:2 (2014). É chegada a hora de a Domínios de Lingu@gem publicar um novo número temático sobre o assunto, acompanhando publicação análoga pela Revista de Estudos da Linguagem 23:3 (2015).

O presente número da Domínios de Lingu@gem, composto por oito artigos, traz trabalhos apresentados durante o XII Encontro de Linguística de *Corpus* (ELC) e da VII Escola Brasileira de Linguística Computacional (EBRALC) sob a temática *Corpus*, Tecnologia e Cultura, além de artigos pertinentes à área recebidos pela revista. Estes eventos ocorreram em novembro de 2014, no Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU). A organização coube aos Profs. Ariel Novodvorsky e Guilherme Fromm, ambos membros do Programa de Pós-Graduação em Estudos Linguísticos (PPGEL/ILEEL/UFU), vinculados ao Grupo de Pesquisas e Estudos em Linguística de *Corpus* – GPELC, da UFU.

Os artigos que compõem este número da Domínios de Lingu@gem cobrem variados focos temáticos da Linguística de *Corpus* e suas interfaces, como será relatado a seguir. As pesquisas discutidas pautam-se no desenvolvimento e utilização de ferramentas computacionais para a análise de dados linguísticos, a utilização da web como *corpus*, criação de atividades didáticas baseadas em *corpora*, incremento de informações pragmáticas extraída de *corpora* a dicionários bilíngues e interfaces da pesquisa terminológica e tradutória com metodologias baseadas em *corpora*.

No primeiro artigo, *Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos*, Diana Santos, Rui Marques, Cláudia Freitas, Alberto Simões e

Cristina Mota apresentam a Gramateca, ambiente para estudos da gramática da língua portuguesa baseados em *corpus*, bem como alguns estudos já desenvolvidos em seu âmbito. O foco desta equipe multiinstitucional é nos procedimentos metodológicos possibilitados pela Gramateca, destacando especialmente o sistema Rêve, que permite revisar e partilhar anotações linguísticas. Dentre os aspectos de destaque apontados no artigo está a possibilidade de cruzamento de anotações humanas com o processamento automático ou semiautomático de dados linguísticos, o que garante o detalhamento qualitativo e a validação quantitativa necessários aos estudos linguísticos contemporâneos.

Milena de Uzeda Garrão, em *A relevância da web como corpus para a identificação de padrões de lexicalização: o caso de “bater + SN” no português brasileiro*, argumenta a favor da utilização da web como recurso empírico para pesquisas linguísticas. A autora explora as inovações de uso do verbo “bater”, notadas tanto no Brasil quanto em Portugal, através das quais esse verbo passa a ter ocorrências transitivas diretas, com sentido de “vencer”, aliadas a ocorrências aparentemente de verbo suporte como “bater um medo”. A pesquisa desenvolvida inicialmente, a partir de textos jornalísticos, levou a autora a exploração ampliada através da web como *corpus* para a sua verificação. Garrão conclui o seu artigo com uma defesa epistemológica favorável ao uso de grandes volumes de dados, como ferramenta necessária a que o pesquisador possa efetivamente contemplar mudanças, ampliações e volatilidades do sistema linguístico em uso.

Em *Extração automática de candidatos a termos do “Curso de Linguística Geral” com apoio de recursos da Linguística de Corpus e do Processamento da Linguagem Natural*, Maria José Borcony Finatto, Lucelene Lopes e Alena Ciulla exploram ferramentas computacionais e comparam sua eficácia para a identificação de sintagmas nominais relevantes ao se tomar a famosa obra póstuma de Ferdinand de Saussure como *corpus* de análise. As pesquisadoras descrevem os passos metodológicos da pesquisa, indicando a anotação morfossintática do *corpus* através do anotador Palavras como elemento um dessa trajetória; a seguir, faz-se a extração de SNs através da ferramenta ExATOlp, a qual faz uso de técnicas estatísticas e linguísticas. Os resultados são consolidados a seguir através de procedimentos de visualização. Os resultados obtidos nesta etapa dos procedimentos são comparados àqueles obtidos com a utilização da ferramenta computacional AntConc, de acesso livre à comunidade. As autoras concluem,

destacando a funcionalidade de ambas as ferramentas testadas, porém destacando o potencial específico dos resultados obtidos via ExATOlp.

Na sequência, Maúcha Andrade Gamonal e Tiago Timponi Torrent apresentam *Diretrizes para a criação de um recurso lexical multilíngue a partir da semântica de frames: a experiência turística em foco*. Neste artigo os autores exploram os passos metodológicos adotados na confecção do Dicionário FrameNet Brasil da Copa do Mundo, um dicionário trilíngue (português-inglês-espanhol) de acesso grátis, lançado para atender às demandas turísticas com o evento da Copa do Mundo celebrada no Brasil em 2014. O projeto, que se beneficia da larga experiência da matriz FrameNet desenvolvida na Universidade de Berkeley, utiliza a Semântica de Frames como marco teórico que orienta suas propostas de entradas no dicionário, orientadas por ocorrências de uso em *corpora*. O artigo apresenta o domínio turístico para ilustrar as funcionalidades da ferramenta descrita.

Maria Cláudia Nunes Delfino aborda em *Designing English teaching activities based on popular music lyrics from a corpus perspective* o potencial das ferramentas disponíveis via Linguística de *Corpus* na composição de atividades que possam fomentar a atuação didática de professores de inglês. A pesquisadora compilou um *corpus* de 150.000 palavras a partir de letras de músicas votadas por seus alunos e o tratou através da ferramenta AntConc. Na sequência, a pesquisadora cotejou os resultados obtidos com listas de palavras mais frequentes no *corpus* de referência de língua inglesa COCA e elaborou atividades aplicadas a seus alunos. A aplicabilidade e impacto do material didático desenvolvido foram testados através de diários reflexivos elaborados pelos alunos, que corroboraram a intuição de que materiais criados a partir de *corpora* propiciam condições favoráveis de aprendizagem.

Ampliando a discussão sobre o estudo terminológico para o domínio bilíngue, Marina Araújo Vieira e Silvana Maria de Jesus, autoras de *O uso de corpora comparáveis na pesquisa terminológica bilíngue*, exploram a interface estudos terminológicos e tradução, via metodologia de exploração de *corpora*. O domínio temático do estudo é o espiritismo, abordado através de obras originalmente escritas em português e traduzidas para o inglês. Termos espíritas, notadamente relativos à mediunidade, foram extraídos via *corpora* bilíngues comparáveis e paralelos. Os resultados levaram à elaboração de fichas terminológicas e seleção de termos multivocabulares a fim de se criar uma amostra de glossário bilíngue. As autoras concluem que a riqueza vocabular oferecida no campo explorado na língua inglesa não foi contemplada nas obras traduzidas através das escolhas

terminológicas dos tradutores, desta forma apontando lacunas a serem preenchidas na formação de tradutores profissionais.

Explorando a interface entre a metodologia da linguística de *corpus* e os estudos do léxico, Renato Railo Ribeiro, no artigo *Ilocuções comissivas em dicionários híbridos italiano>português brasileiro: proposta de dicionarização a partir do uso de corpora*, investiga o enriquecimento de verbetes em dicionários bilíngues através de inserção de informações ilocucionárias em ambas as línguas representadas. Para tal fim, o autor apresenta os critérios adotados para a exploração de valores ilocucionários em *corpora*, indicando a pesquisa exploratória a partir de cinco verbos em ambas as línguas enfocadas, e como se daria a inserção de tais informações em dicionários bilíngues. O autor conclui apontando os ganhos que a introdução de dados pragmáticos em verbetes de dicionários bilíngues poderiam acarretar ao aprendiz que deles faz uso.

Também explorando a interface Linguística de *Corpus* e Estudos da Tradução, o artigo que fecha este número da Domínios de Lingu@gem, *Tradução e Retradução de “The Picture of Dorian Gray”, de Oscar Wilde: um estudo de corpus com foco na apresentação do discurso*, de Líbia Cremonez Domingos e Igor A. Lourenço da Silva, investiga um *corpus* compilado a partir de trechos da obra original “The Picture of Dorian Gray”, de sua primeira tradução para o português e duas retraduições. O propósito dos autores é avaliar a hipótese de que a primeira tradução de uma obra literária é “incompleta e domesticadora”. A partir da metodologia instaurada, que utilizou-se de etiquetagem do *corpus*, seu alinhamento e processamento semi-automático, os autores concluem que há indícios que corroboram a hipótese investigada.

Esperamos que este número temático da Domínios de Lingu@gem auxilie a divulgação da pesquisa que vem sendo desenvolvida no campo temático da Linguística de *Corpus*, bem como estimule a curiosidade de pesquisadores ainda não familiarizados com as metodologias empregadas nos estudos aqui relatados na busca de interfaces com seus próprios interesses de pesquisa.

Heliana Mello (UFMG)

Referências

Letras & Letras, 30:2, 2014. Disponível em: <http://www.seer.ufu.br/index.php/letraseletras/issue/view/1217>. Acesso em 17/12/2015.

FROMM, G.; NOVODVORSKY, A. Triangulando *corpus*, tecnologia e cultura: ELC e EBRALC na UFU. Apresentação. **Revista de Estudos da Linguagem**. 23:3, 2015, no prelo.

Revista Brasileira de Linguística Aplicada, 11:2, 2011. Disponível em : http://www.scielo.br/scielo.php?script=sci_issuetoc&pid=1984639820110002&lng=es&nrm=1. Acesso em 17/12/2015.

Revista de Estudos da Linguagem 23:3, 2015, no prelo. Disponível em: <http://www.periodicos.letras.ufmg.br/index.php/relin>. Acesso em 17/12/2015.

Veredas – Revista de Estudos Linguísticos, 13:2, 2009. Disponível em : <http://www.ufjf.br/revistaveredas/edicoes/2009-3/2009-2/>. Acesso em 17/12/2015.

Domínios de Lingu@gem