DOI: 10.17851/2237-2083.23.3.589-597

Triangulando *corpus*, tecnologia e cultura: ELC e EBRALC na UFU

Este número da *Revista de Estudos da Linguagem* (RELIN) está voltado para a metodologia e a abordagem baseadas em *corpora*, e leva até o leitor textos oriundos de trabalhos advindos do XII Encontro de Linguística de Corpus (ELC) e da VII Escola Brasileira de Linguística Computacional (EBRALC). Os eventos foram realizados nas dependências do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU), Minas Gerais, em novembro de 2014. Esses eventos contaram com o patrocínio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES –, da Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG – e do Programa de Pós-Graduação em Estudos Linguísticos – PPGEL/ILEEL/UFU. A comissão organizadora foi composta por professores da própria instituição, da Universidade Federal de Minas Gerais – UFMG, da Universidade Federal do Rio Grande do Sul – UFRGS e da Universidade de São Paulo – USP, que contou com a participação ativa dos estudantes vinculados ao nosso Grupo de Pesquisas e Estudos em Linguística de *Corpus* – GPELC,[1] da UFU.

Nessa ocasião e com o tema *Corpus, tecnologia e cultura*, reunimos aproximadamente cento e dez participantes, que apresentaram seus trabalhos em diferentes modalidades, a saber: pôsteres, o já consagrado minuto de loucura e comunicações orais. Além de professores convidados de diferentes instituições nacionais, que plasmaram sua contribuição em palestras, mesa-redonda e oficinas, pudemos materializar a participação de dois professores de instituições internacionais, Carlos Ramisch (Universidade de Marseille, França) e Giovanni Parodi (PUC-Valparaíso, Chile), que compartilharam em plenária seus conhecimentos, pesquisas e recursos da área desenvolvidos no exterior.

[1] Disponível em: <http://dgp.cnpq.br/dgp/espelhogrupo/0919828121533301>. Acesso em: 19 nov. 2015.

No Brasil, apesar dos avanços consideráveis da última década, os trabalhos e resultados de pesquisa em Linguística de *Corpus* e Linguística Computacional ainda não contam com a merecida difusão. Um dos nossos objetivos foi o de mostrar os benefícios para a sociedade brasileira, pela promoção do diálogo entre as áreas da Linguística e da Computação tanto pela divulgação de produtos concretos desenvolvidos para o desempenho de tarefas imediatas como pelo subsídio para a realização de pesquisas e consequente descrição e análise linguísticas. Buscando oportunizar uma experiência nas mais diversas áreas envolvidas, durante os dias de encontro em Uberlândia, promovemos a familiarização com conceitos, abordagens e metodologias relacionadas à compilação, tratamento, etiquetagem e exploração de *corpora*, para as mais diferentes aplicações.

O ELC e a EBRALC são eventos itinerantes Brasil afora, que reúnem pesquisadores brasileiros e convidados internacionais, para a discussão de pesquisas de base empírica. Como bem pontuado por Novodvorski e Finatto (2014), o desenvolvimento da Linguística de *Corpus* (LC) no país já se consolida como uma aventura mais do que adequada. Desde a publicação do livro pioneiro de Berber Sardinha (2004), já são doze anos de um desenvolvimento contínuo de inúmeras pesquisas, que estão revolucionando os objetos de estudo das mais de quarenta subáreas da Linguística (FROMM; YAMAMOTO, 2013). Utilizando-se de dados concretos, baseados em *corpora* cuidadosamente compilados, preparados e, por vezes, também etiquetados, constatamos uma mudança de foco nas análises linguísticas brasileiras, especialmente naquelas que se baseiam na intuição dos pesquisadores como ponto de partida, levando-as para um patamar muito próximo das ciências exatas. Por outro lado, também cabe mencionar que inúmeros programas de análise lexical, como o *WordSmith Tools*, versão 6.0 (SCOTT, 2012)[2] e o *AntConc*, versão 3.4.4. (ANTHONY, 2014),[3] muitas das vezes porta de entrada para este contexto e vários outros que dão suporte a pesquisas

[2] Disponível em: <http://www.lexically.net/wordsmith/downloads/>. Acesso em: 10 nov. 2015.

[3] Disponível em: <http://www.laurenceanthony.net/software/antconc/>. Acesso em: 10 nov. 2015.

mais pontuais, como *UAM Corpus Tool, 3.2j* (O'DONNELL, 2015)[4] e *Sketch Engine* (KILGARRIFF *et al*., 2014),[5] estão sendo empregados em trabalhos dos mais diferentes campos do saber.

Os trabalhos na área da LC já foram tema de publicação em diferentes revistas dedicadas à Linguística: *Domínios de Lingu@gem*[6] e *Letras & Letras*[7] (UFU) e *Veredas*[8] (UFJF). Chegou o momento de a RELIN apresentar sua contribuição à área. Por meio de trabalhos baseados e / ou direcionados por *corpora* (TAGNIN, 2015), apresentamos aqui um pequeno panorama do que está sendo realizado em pesquisas linguísticas de ponta, de norte a sul do país.

Pablo Neves Machado e Vera Lúcia Strube de Lima, pesquisadores da PUC-RS, exploram em detalhe, no seu texto *Extração de relações hiponímicas em um corpus de língua portuguesa*, a importância das relações hiponímicas, na construção de estruturas de conhecimento (ontologias ou taxonomias), num *corpus* de língua portuguesa. Os autores objetivam o aprimoramento dos procedimentos de busca e extração de padrões, pela alimentação de um protótipo que gera as relações hiponímicas, e o processo de avaliação humana na produtividade dos padrões resultantes. As principais contribuições apontadas no trabalho dizem respeito à integração, em uma mesma pesquisa, de regras propostas por diferentes pesquisadores no assunto e pela análise detalhada dos resultados.

Com o texto *O léxico do corpo e anotação de sentidos em grandes corpora – o projeto Esqueleto*, Cláudia Freitas, Bruno Carriço (ambos da PUC-Rio), Diana Santos, Heidi Jansen (ambas da Universidade de Oslo) e Cristina Mota (Linguateca) mergulham num estudo sobre o léxico do corpo humano, motivado pelo sentido, com base

---

4 Disponível em: <http://www.wagsoft.com/CorpusTool/index.html>. Acesso em: 10 nov. 2015.

5 Disponível em: <https://www.sketchengine.co.uk/>. Acesso em: 15 nov. 2015.

6 Disponível em: <http://www.seer.ufu.br/index.php/dominiosdelinguagem/issue/view/615>. Acesso em: 19 nov. 2015.

7 Disponível em: <http://www.seer.ufu.br/index.php/letraseletras/issue/view/1217>. Acesso em: 19 nov. 2015.

8 Disponível em: <http://www.ufjf.br/revistaveredas/edicoes/2009-3/2009-2>. Acesso em: 19 nov. 2015.

em anotação e revisão de *corpora* de grande extensão em língua portuguesa. Os autores destacam a relevância do processo de anotação, caracterizado como "forma de estudo", que operaria "como uma lente de aumento", aliado ao processamento automático da língua e de sua descrição. Para além do detalhamento dos procedimentos metodológicos, os autores explicitam as diferentes informações pertinentes aos processos de anotação, às classes compreendidas e à exploração do *corpus* compilado de entrevistas e textos literários, que possui mais de 2,5 milhões de palavras. Uma informação relevante para o leitor é que todo o material está disponibilizado para consulta na página do projeto e, ainda, são oferecidas diversas sugestões para a realização de novas pesquisas, explorando os diversos sentidos do corpo na língua portuguesa.

Em *A tradução de suculentos jogos de palavras, sem perder o sabor*, Stella Tagnin (USP) concentra sua atenção em diferenças culturais entre a Inglaterra e o Brasil, pelo estudo da tradução para o português, no âmbito da culinária. A pesquisadora analisa, especificamente, a tradução dos títulos de receitas de um livro britânico de sucos para crianças. Segundo Tagnin, a maioria dos títulos analisados apresentam jogos de palavras e usos metafóricos, que exigem a busca de soluções por parte do tradutor, no sentido de preservar a riqueza semântica presente no apelo para o público infantil, realizado por meio de diferentes recursos como colocações, aliterações, rimas e referências culturais. Considerando as estratégias empregadas para a denominação dos sucos na língua de chegada e também observando a priorização do 'efeito' sobre a 'forma', em termos funcionalistas, a pesquisadora aponta que tais objetivos foram alcançados nas traduções e que, quando necessário, foram feitas adaptações culturais, deixando resguardado o caráter apelativo.

A proposta do texto *Topic Modeling for Keyword-Phrase Extraction: using Natural Language Processing methods for keyword extraction in Portal Min@s*, de Arnaldo Candido Junior (UTFPR), Célia Magalhães (UFMG), Helena Caseli (UFSCar) e Régis Zangirolami (USP - São Carlos), é apresentar uma comparação de métodos, baseados na Linguística de Corpus e no Processamento de Linguagem Natural, para extração de palavras-chave em textos ficcionais (romances, mais especificamente). Duas ferramentas, cujas técnicas de extração são

diferentes, foram analisadas no trabalho: uma mais clássica e comum aos linguistas, o *WordSmith Tools*, e outra com enfoque diferente, baseada em tópicos, o *Latent Dirichlet Allocation* (LDA), que pode ser acessada pelo *site Portal Min@s: Corpora de Fala e Escrita*. O *corpus* de análise consiste em três diferentes traduções para o português do romance *Heart of Darkness*, de Joseph Conrad. Os autores concluem que, no atual estágio de desenvolvimento de programas voltados para a análise linguística das palavras-chave, o uso do *WordSmith Tools* ainda seria o mais apropriado.

Comparação linguística e perfilação gramatical sistêmica em um corpus combinado, de Francieli Silvéria Oliveira (UFOP), analisa a organização gramatical e semântica, no par linguístico inglês / português brasileiro, de um *corpus* de manuais de instrução. Entre os objetivos principais do artigo, a autora aponta o estudo da variação linguística do registro, a comparação dos sistemas linguísticos envolvidos e a descrição da produção textual das traduções. Os padrões identificados por meio da análise, levaram a autora à afirmação de que o manual de instrução teria a macro função de 'capacitar' para o desenvolvimento de determinada atividade, segundo instruções, e não pela regulação de comportamentos. As funções semânticas de 'explicar', 'classificar', 'introduzir' e 'comandar' comporiam, desse modo, a macro função de 'capacitar'. Para além das semelhanças, a pesquisadora também observou diferenças no par linguístico estudado, no que diz respeito a funções semânticas como 'convidar', presente apenas nos manuais originais em língua inglesa, e 'introduzir' como Processo Verbal, somente observado no *subcorpus* de português original. Oliveira destaca que o método de pesquisa empregado, que integra a perfilação sistêmica e a comparação linguística, abre novas perspectivas na análise de *corpus*.

O texto *The relevance of the Sketch Engine software to build Field – Football Expressions Dictionary*, de Rove Luiza de Oliveira Chishman, Aline Nardes dos Santos, Diego Spader de Souza e João Gabriel Padilha (todos da UNISINOS), tece considerações acerca dos recursos que oferece a ferramenta *Sketch Engine* e de sua potencialidade para a análise de *Frames* Semânticos no desenvolvimento de um dicionário trilíngue de expressões de futebol, baseado em *corpora* linguísticos. Os pesquisadores destacam que um dos recursos do

programa, o *Word Sketch*, mostra o comportamento sintático-semântico do *frame*, por meio de evidências empíricas. Após a abrangência da fundamentação teórica e do potencial metodológico da Linguística de *Corpus*, os autores descrevem os procedimentos de análise, aliados empiricamente à identificação das unidades polissêmicas e das colocações presentes no *corpus*, fundamentalmente com o recurso *Word Sketch*, que possibilita a visualização detalhada dos usos produtivos da língua.

Marco A. Sobrevilla Cabezudo, Erick G. Maziero, Márcio S. Dias, Paula C. F. Cardoso, Pedro P. Balage Filho, Thiago A. S. Pardo, Verônica Agostini, Fernando A. A. Nobrega (todos do NILC/USP - S. Carlos), Jackson W. C. Souza e Ariani Di Felippo (ambos da UFSCar) e Cláudia D. Barros (IFET - São Paulo), em seu texto denominado *Anotação de Sentidos de Verbos no corpus CSTNew*, abordam um dos temas mais complexos para o Processamento de Linguagem Natural (PLN): a ambiguidade lexical. Considerando os ambientes de ocorrência, a polissemia das palavras impõe um desafio para o PLN, a determinação do sentido adequado de uma palavra em contexto. Conforme explicam os autores, a responsável por essa tarefa em PLN é a Desambiguação Lexical de Sentido (DLS). Para a realização de trabalhos dessa natureza, tanto o processo e desenvolvimento de métodos de anotação como o uso do *corpus* anotado é fundamental, para uma análise mais profunda da ambiguidade, isto é, a consulta a um *corpus* com anotação semântica dos verbos expande as possibilidades da DLS. Nesse sentido, o trabalho apresentado na publicação descreve os procedimentos pertinentes à anotação de um *corpus* jornalístico, por meio da *WordNet* de Princeton, como repositório de sentidos. Para além do detalhamento dos processos de etiquetagem e extração de dados e obtenção de resultados, uma das principais contribuições deste texto, resultante do projeto denominado SUCINTO, é a disponibilização para consulta do *corpus* jornalístico anotado e da própria ferramenta de edição, no intuito de poder subsidiar pesquisas futuras em DLS, tal como expressam seus autores.

No texto intitulado *Collocations workbook: um material de apoio pedagógico on-line e baseado em corpus para o ensino de colocações em inglês*, Adriane Orenha-Ottaiano (UNESP) justifica a elaboração de materiais de ensino que contemplem "uma seleção cuidadosa de

colocações", ajustada às principais dificuldades e necessidades dos aprendizes de inglês como língua estrangeira, principalmente dirigida a falantes de português brasileiro. Desse modo, a autora aponta ao tratamento das colocações nas aulas de inglês, buscando preencher a carência de propostas fraseológicas para o ensino, por meio da compilação de um *Workbook online* de colocações. Alimentam esse trabalho as produções de alunos universitários de tradução que, por sua vez, são parte integrante de um *corpus* paralelo de aprendizes, na direção português / inglês, além de redações que também compõem esse banco de dados. Por meio de uma combinação entre o programa *WordSmith Tools* e a consulta ao *Corpus of Contemporary American English*, é constatada a frequência dos padrões colocacionais. Com o objetivo de promover o desenvolvimento da fluência em língua inglesa, a pesquisadora destaca a importância de buscar conscientizar quanto ao fenômeno da colocação. Outro fator relevante da pesquisa é a disponibilização do recurso na *Web*.

Por último, Tommaso Raso (UFMG), Maryualê Malvessi Mittmann (FACVEST) e Anna Carolina Oliveira Mendes (UFMG) analisam a pausa silenciosa, como critério de segmentação da fala em unidades comunicativamente autônomas. Em seu texto intitulado *O papel da pausa na segmentação prosódica de corpora de fala*, os autores contextualizam e discutem as noções de unidade de referência da fala e de pausa. As fronteiras prosódicas delimitariam em unidades pragmáticas a segmentação da fala. Isto é, pelo estudo da fala espontânea com base no *corpus* C-ORAL-BRASIL e, neste caso particular, pelo estudo da duração das pausas, a investigação procurou estabelecer critérios confiáveis para a identificação de fronteiras de unidades de referência. Contudo, os pesquisadores chegam à conclusão acerca da impossibilidade de vincular genericamente as marcas físicas de pausas, o silêncio, no caso, às fronteiras de unidade de referência da fala. O estudo da quebra prosódica, em contrapartida, tem-se mostrado mais profícuo para a finalidade da segmentação da fala espontânea. Na expectativa de compreender melhor a percepção da natureza da fala, os autores adotam uma abordagem guiada pelo *corpus* e segundo os princípios de aprendizagem de máquina, reconhecendo a grandiosidade de desenvolver

um sistema (semi) automatizado para a segmentação de enunciados e de unidades tonais da fala.

Esperamos que todos tenham um bom proveito acadêmico com os textos aqui apresentados, resultantes do nosso encontro no triângulo mineiro, por ocasião do XII ELC e da VII EBRALC realizados na UFU, e que esses textos nos inspirem a continuar investindo nas pesquisas empíricas, para as quais a Linguística de *Corpus* vem demonstrando ser uma das principais aliadas no século 21.

Ariel Novodvorski (UFU)
Guilherme Fromm (UFU)

Referências

ANTHONY, L. *AntConc* (Version 3.4.4). Tokyo: Waseda University, 2014. Disponível em: <http://www.laurenceanthony.net/software.html>. Acesso em: 15 nov. 2015.

BERBER SARDINHA, T. *Lingüística de Corpus*. Barueri S.P.: Editora Manole, 2004. 410 p.

FROMM, G.; YAMAMOTTO, M. Terminologia, Terminografia, Tradução e Linguística de *Corpus*. In: TAGNIN, S. E. O.; BEVILACQUA, C. (Org.). *Corpora na Terminologia.* São Paulo: Hub Editorial, 2013. p.129-151.

NOVODVORSKI, A.; FINATTO, M. J. B. Linguística de corpus no Brasil: uma aventura mais do que adequada. *Letras & Letras*, v. 30, n. 2, Jul/dez 2014. Disponível em: <http://www.seer.ufu.br/index.php/letraseletras/article/view/28516/15799>. Acesso em: 14 nov. 2015.

SCOTT, M. *WordSmith Tools*. Versão 6, 2012. Disponível em: <http://www.lexically.net/wordsmith/downloads>. Acesso em: 15 nov. 2015.

TAGNIN, S. E. O. Glossário de Linguística de Corpus. In: VIANA, V.; TAGNIN, S. E. O. (Org.). *Corpora na tradução*. São Paulo: Hub Editorial, 2015. 331p.