ANOTAÇÃO SEMÂNTICA E MORFOSSINTÁTICA PARA IDENTIFICAÇÃO DE PALAVRAS E *CHUNKS* EM *CORPORA* TEXTUAIS EM LÍNGUA PORTUGUESA: PROBLEMATIZAÇÃO E PROPOSTA PRELIMINAR

Luiz Henrique Santana Neves
Universidade de Brasília-UnB
Teoria e análise linguística: Léxico e Terminologia
1º/2011- 2º/2014

A delimitação da unidade palavra constitui tema de amplo debate no âmbito dos estudos teóricos da linguagem. Não poucos pesquisadores têm apresentado fatos que evidenciam o caráter intuitivo, variável e fluído dessas unidades teóricas de análise. Como consequência, não se conseguiu, ainda, estabelecer uma definição precisa e universal para o termo, dados, entre outros aspectos, os diferentes tipos linguísticos existentes (BIDERMAN, 1999; HENGEVELD, 2003). Essas questões repercutem diretamente em aplicações linguísticas que buscam, por exemplo, o rigor científico na descrição, comparação, contraste e/ou processamento da linguagem. Os resultados obtidos, por exemplo, na busca de palavras-chave e de *chunks* 'combinações ou bloco de palavras', em extensos *corpora* eletrônicos, etiquetados morfossintaticamente, têm se mostrado, muitas vezes, insatisfatórios. Ao priorizar, somente, critérios linguístico-formais e estatísticos no tratamento dos vocábulos textuais em *corpora* desse tipo, deixa-se de contemplar a linguagem no seu aspecto integral ou holístico. Por conseguinte, eludem-se, entre outras, análises e marcações semânticas e/ou pragmático-discursivas, que propiciariam resultados de busca mais factuais e precisos em relação, não somente, aos *chunks*, mas também aos vocábulos unitários ou discretos. Dessa maneira, objetiva-se, nesta exposição, apresentar uma proposta teórico-metodológica, preliminar, para a anotação semântica e morfossintática de vocábulos, com vistas à obtenção de respostas mais efetivas, na busca de unidades fraseológicas ou *chunks*, em extensos *corpora* eletrônicos, na variedade brasileira da língua portuguesa. Para tanto, priorizar-se-á os fatos de língua concernentes às combinações de palavras, designadas construções com verbos-suporte. Tendo em vista os resultados alcançados em pesquisas preliminares, destaca-se o papel linguístico, desempenhado pelos substantivos que ocupam a posição de objeto direto dos verbos-suporte. Partindo, então, da análise de dados empíricos, extraídos de um *corpus* composto de 33 milhões de *tokens*, no Português do Brasil, propõe-se a subsequente metodologia para alcançar o citado objetivo: (i) que se adote a seguinte definição de substantivos abstratos: substantivos que "não têm referente no mundo, mas constituem-se em atos, eventos, estados relacionados a seres ou estado de coisas (corrida, lverdeza, crença, etc.)" (BORBA, 1996, p. 65); (ii) que, em todos os contextos nos quais ocorram os verbos fazer, dar, ter e tirar em combinação com substantivos, em posição de objeto direto, distingam-se substantivos concretos de "abstratos": dar uma corrida < substantivo abstrato, dar uma jóia < substantivo concreto; (iii) que, nas combinatórias com os verbos fazer, dar, ter e tirar, em que se identifiquem substantivos abstratos, em posição de objeto direto, marquem-se os mesmos com a etiqueta semântica Subst.Abstr., e os referidos verbos com a etiqueta V^{sup}: fazer_{V^{sup}}. força_{Subst.Abstr.}, tirar_{V^{sup}}. proveito_{Subst.Abstr.}; (iv) que as combinações de verbos-suporte e substantivos abstratos sejam identificadas como um bloco linguístico-informacional, marcado com a etiqueta categorial Construção com verbos-suporte (C. v_{sup}); (v) que as construções com verbos-suporte se identifiquem, morfossintaticamente, por meio da marca V, correspondente à classe gramatical verbo, dado o valor de categoria verbal, denotado por esses *chunks*, nos contextos oracionais

nos quais figuram. Espera-se que o método apresentado possa contribuir para uma melhor identificação de vocábulos discretos, e principalmente, de fraseologismos no processamento da informação textual, para constituição de extensos *corpora*.

Referências bibliográficas

BIDERMAN, M. T. O conceito linguístico de palavra. **Palavra**, n. 5. Rio de Janeiro: Departamento de letras da PUC. 1999. 81-97 p.

BORBA, F. **Uma Gramática de Valências para o Português**. São Paulo: Ática. 1996.

HENGEVELD, K. La tipología lingüística. **En torno a los universales lingüísticos**. Cambridge University press- AKAL. [S.I.], 2003. 89-111 p.

LEWIS, M. **The lexical approach**. Londres. Language Teaching Publications. 1993.