Procura-PALavras (P-PAL): Uma aplicação web para uma base de dados lexical do português europeu. [*]

Álvaro Iriarte S., Ana Paula Soares, Alberto Simões, José João Almeida, Montserrat Comesaña, Ana Costa, João Machado, Patrícia França.

27 Congrés International de Linguistique et de Philologie Romanes,
Section 16. Projets en cours ; ressources et outils nouveaux

1. Introdução

Neste trabalho apresentamos o projeto Procura-PALvras (P-PAL) cujo principal objetivo é desenvolver uma aplicação web que permite a obtenção de uma grande diversidade de métricas lexicais e sublexicais para ≈209.000 formas e ≈52.000 lemas do português europeu (PE) contemporâneo extraídas de um corpus de grandes dimensões (superior a 227 milhões de palavras).

O P-PAL é mais do que uma versão adaptada para o PE do software inglês *N-Watch* (Davis, 2005), já adaptado para a língua espanhola como *BuscaPalabras* (B-Pal) (Davis & Perea, 2005). A nossa versão pretende ser um instrumento de âmbito multidisciplinar e que sirva múltiplos objetivos e diferentes áreas de pesquisa (linguística, psicolinguística, processamento de linguagem natural, etc.), incluindo, para isso, outros índices não contemplados nos seus congéneres e uma dupla possibilidade de utilização: (i) obter palavras que obedeçam a determinados requisitos; ou (ii) analisar palavras num conjunto requisitos.

[*] O projeto P_PAL (PTDC/PSI-PCO/104679/2008) é financiado pela FCT (Fundação para a Ciência e a Tecnologia), pelo QREN (Quadro de Referência Estratégica Nacional) e pelo COMPETE (Programa Operacional Factores de Competitividade), um programa criado pela União Europeia como parte do Fundo Europeu de Desenvolvimento (FEDER).

O P-PAL permitirá, para além da computação por defeito do valor de frequência de todas as suas entradas lexicais a realização de um conjunto diversificado de análises relativas quer às dimensões morfossintáticas (categoria gramatical, frequência de lema, etc.) e ortográficas (número de letras, estrutura consoante-vogal, etc.); quer às dimensões fonético-fonológicas (pronunciação da palavra, número de fonemas, diversas medidas de frequências de bifone, etc.), silábicas (silibificação frequências de tipo silábico e *token,* etc.), e de vizinhança (vizinhos por substituição, adição, subtração ou transposição, etc.). Incluirá também índices léxico-semânticos ausentes em qualquer base lexical portuguesa (listas de colocações e coocorrentes frequentes; hiperónimos e hipónimos, etc.). Permitirá ainda a análise dos valores normativos para os parâmetros subjetivos de imaginabilidade, concreteza, familiaridade, valência, ativação e controlabilidade, de interesse para os estudos na área da psicolinguística e ainda não disponíveis entre nós ou disponíveis para um léxico bastante restrito.

A aplicação, aberta e de acesso livre, pode ser consultada em http://p-pal.di.uminho.pt/tools.

Bibliografia

- Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods, 37(1),* 65-70.

- Davis C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, *37,* 665-671.

- Mateus, M. H. M., Brito, A. M., Duarte, I., Faria, I. H. et al. (2003). *Gramática da Língua Portuguesa* (5ª ed.). Lisboa: Editorial Caminho.

- Rocha, P., & Santos, D. (2000). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In M. G. V. Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)* (pp. 131-140). São Paulo, Brasil.

Soares, A. P., Comesaña, M., Iriarte, A., Almeida, J. J., Simões, A., Costa, A., França, P. C., & Machado, J. (2010). P-PAL: Uma base lexical com índices psicolinguísticos do Português Europeu. *Linguamática, 2(3)*, 67-72.