Procura-PALavras (P-PAL): A web application for a new European Portuguese lexical database

Ana Paula Soares1, Montserrat Comesaña1, Álvaro Iriarte2, José João de Almeida3, Alberto Simões3, Ana Costa3, Patricia França1 & João Machado1

1Psycholinguistics Research Group, 2Centre for Humanistic Studies, 3Computer Science and Technology Center, University of Minho

Procura-PALavras (P-PAL) is a web application for a new European Portuguese (EP) lexical database that provides a series of objective (lexical and sublexical) and subjective indices for around 160,000 non-lemmatized and around 420,000 lemmatized EP words. Based on a corpus of over 225 million EP words, the P-PAL enables users to obtain a broad range of statistics concerning the properties of word stimuli, including several indices of word frequency, syllable frequency, bigram and trigram frequency, orthographic and phonological structure, morphological structure, semantic structure, subjective, and objective characteristics; orthographic and phonological similarity, lexical semantic indexes, concreteness, familiarity, imageability, valence, arousal, and dominance measures. In order to obtain these statistics the user should decide between a lemma or wordform search in the application and between two word-based queries: (i) analyze word lists in specific characteristics (objective and/or subjective); or (ii) generate lists of words with specific characteristics (objective and/or subjective). In this work we present the lemma and wordform frequency indices available, as well as some structure and similarity orthographic measures such as word length, neighborhood density and frequency, letters and bigram type and token frequencies are also presented.