

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

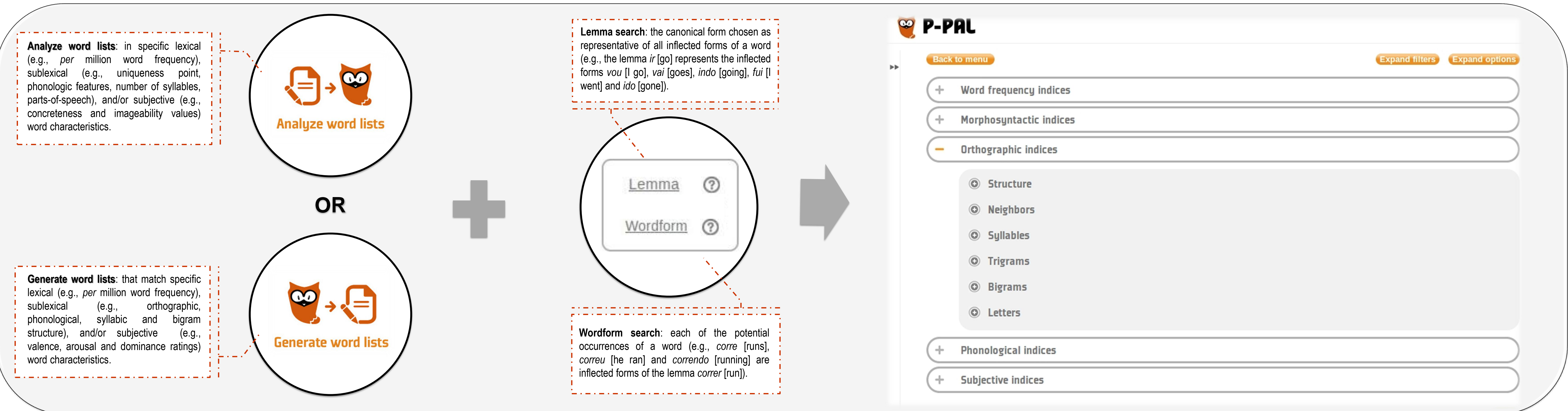
Ana Paula Soares<sup>1</sup>, Álvaro Iriarte<sup>2</sup>, José João de Almeida<sup>3</sup>, Alberto Simões<sup>2,3</sup>, Manuel Perea<sup>4</sup>, Ana Costa<sup>1</sup>, Patrícia França<sup>1</sup>, João Machado<sup>1</sup>, Montserrat Comesaña<sup>1</sup>, & Andreia Rauber<sup>5</sup>

<sup>1</sup>Escola de Psicologia, <sup>2</sup>Instituto de Letras e Ciências Humanas, <sup>3</sup>Departamento de Informática, Universidade do Minho, Portugal

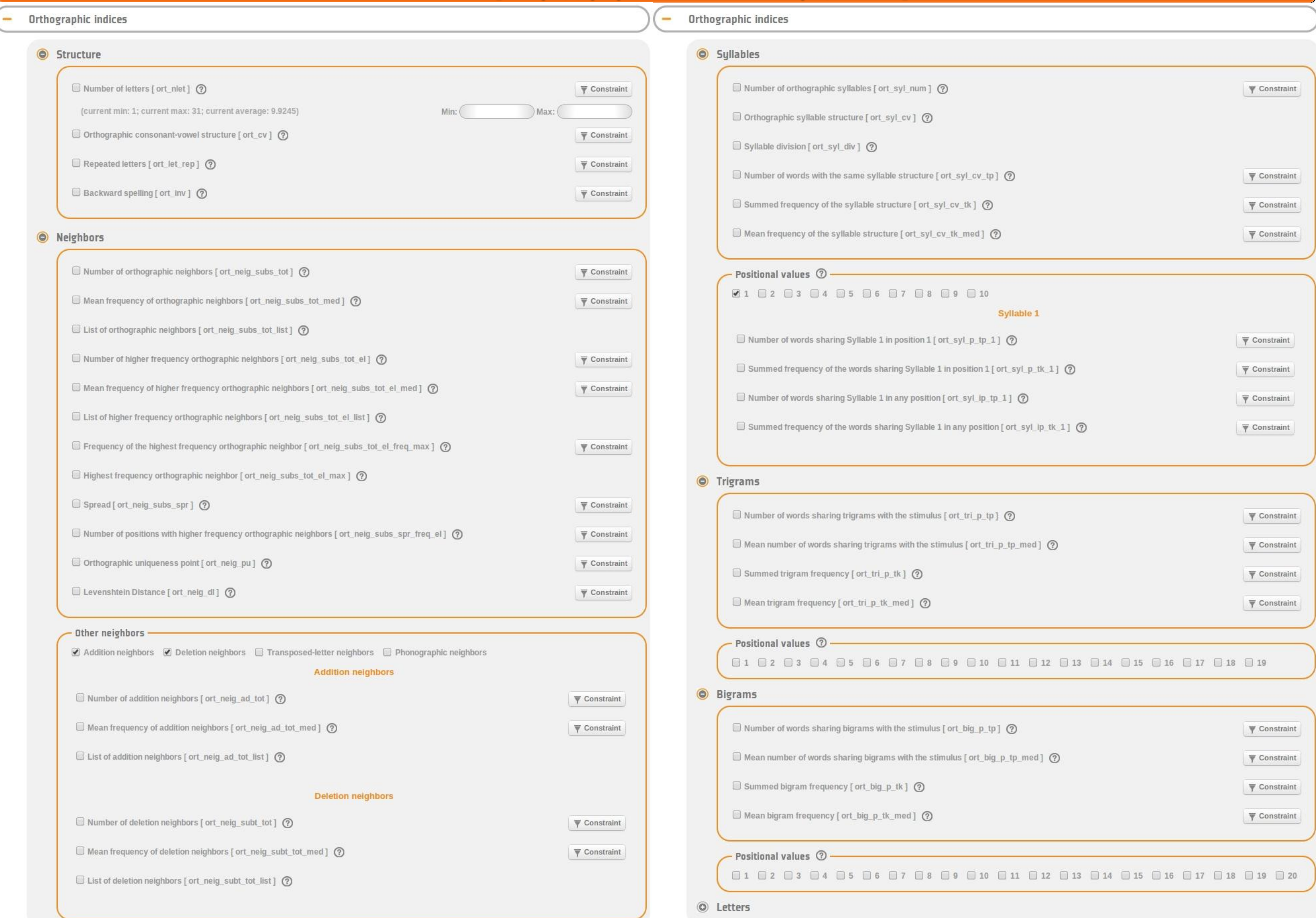
<sup>4</sup>Departamento de Metodología, Universitat de València, Spain, & <sup>5</sup>Universidade Católica de Pelotas, Brazil

Language and Neuroscience Conference | Federal University of Santa Catarina | Florianópolis, Brazil | November 29 - December 1, 2012

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) word statistics for  $\approx 210,000$  non-lemmatized and  $\approx 54,000$  lemmatized EP words. Based on a *corpus* of over 200 million EP words, P-PAL enables users to obtain a broad range of indices, including several measures of word frequency (e.g., *per* million words,  $\log_{10}$  frequencies), morphosyntactic information, orthographic (e.g., number of letters, orthographic uniqueness point, bigrams), phonological (e.g., IPA phonetic transcription, number of phones, biphones), syllable (e.g., orthographic and phonological syllabification), and neighborhood (e.g., substitution, addition, deletion and transposition-letter neighbors) measures. P-PAL also offers norms for subjective indices of imageability, concreteness, subjective frequency, valence, arousal and dominance for 3,800 EP words. 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; or (ii) generate lists of words with specific characteristics. P-PAL is a valuable and useful tool for research areas such as Psycholinguistics, Neuroscience, Cognitive Psychology, and Linguistics.



## P-PAL [<http://p-pal.di.uminho.pt/tools>]



**Orthographic indices**

- Structure**
  - Number of letters [ort\_nlet] (current min: 1; current max: 31; current average: 9.9245)
  - Orthographic consonant-vowel structure [ort\_cv]
  - Repeated letters [ort\_let\_rep]
  - Backward spelling [ort\_inv]
- Neighbors**
  - Number of orthographic neighbors [ort\_neig\_subs\_tot]
  - Mean frequency of orthographic neighbors [ort\_neig\_subs\_tot\_med]
  - List of orthographic neighbors [ort\_neig\_subs\_tot\_list]
  - Number of higher frequency orthographic neighbors [ort\_neig\_subs\_tot\_el]
  - Mean frequency of higher frequency orthographic neighbors [ort\_neig\_subs\_tot\_el\_med]
  - List of higher frequency orthographic neighbors [ort\_neig\_subs\_tot\_el\_list]
  - Frequency of the highest frequency orthographic neighbor [ort\_neig\_subs\_tot\_el\_freq\_max]
  - Highest frequency orthographic neighbor [ort\_neig\_subs\_tot\_el\_max]
  - Spread [ort\_neig\_subs\_spr]
  - Number of positions with higher frequency orthographic neighbors [ort\_neig\_subs\_spr\_freq\_el]
  - Orthographic uniqueness point [ort\_neig\_pu]
  - Levenshtein Distance [ort\_neig\_dl]
- Other neighbors**
  - Addition neighbors**
    - Number of addition neighbors [ort\_neig\_ad\_tot]
    - Mean frequency of addition neighbors [ort\_neig\_ad\_tot\_med]
    - List of addition neighbors [ort\_neig\_ad\_tot\_list]
  - Deletion neighbors**
    - Number of deletion neighbors [ort\_neig\_subt\_tot]
    - Mean frequency of deletion neighbors [ort\_neig\_subt\_tot\_med]
    - List of deletion neighbors [ort\_neig\_subt\_tot\_list]

**Orthographic indices**

- Syllables**
  - Number of orthographic syllables [ort\_syl\_num]
  - Orthographic syllable structure [ort\_syl\_cv]
  - Syllable division [ort\_syl\_div]
  - Number of words with the same syllable structure [ort\_syl\_cv\_tp]
  - Summed frequency of the syllable structure [ort\_syl\_cv\_tk]
  - Mean frequency of the syllable structure [ort\_syl\_cv\_tk\_med]
- Positional values**
  - Syllable 1
    - Number of words sharing Syllable 1 in position 1 [ort\_syl\_p\_tp\_1]
    - Summed frequency of the words sharing Syllable 1 in position 1 [ort\_syl\_p\_tk\_1]
    - Number of words sharing Syllable 1 in any position [ort\_syl\_ip\_tp\_1]
    - Summed frequency of the words sharing Syllable 1 in any position [ort\_syl\_ip\_tk\_1]
- Trigrams**
  - Number of words sharing trigrams with the stimulus [ort\_tri\_p\_tp]
  - Mean number of words sharing trigrams with the stimulus [ort\_tri\_p\_tp\_med]
  - Summed trigram frequency [ort\_tri\_p\_tk]
  - Mean trigram frequency [ort\_tri\_p\_tk\_med]
- Positional values**
  - 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
- Bigrams**
  - Number of words sharing bigrams with the stimulus [ort\_big\_p\_tp]
  - Mean number of words sharing bigrams with the stimulus [ort\_big\_p\_tp\_med]
  - Summed bigram frequency [ort\_big\_p\_tk]
  - Mean bigram frequency [ort\_big\_p\_tk\_med]
- Positional values**
  - 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
- Letters**