Behavior Research Methods
https://doi.org/10.3758/s13428-018-1058-z

Procura-PALavras (P-PAL): A Web-based interface for a new European Portuguese lexical database

Ana Paula Soares[1] · Álvaro Iriarte[2] · José João de Almeida[3] · Alberto Simões[2,3] · Ana Costa[1] · João Machado[1] · Patrícia França[1] · Montserrat Comesaña[1] · Andreia Rauber[1,4] · Anabela Rato[1,5] · Manuel Perea[6]

Abstract

In this article, we present Procura-PALavras (P-PAL), a Web-based interface for a new European Portuguese (EP) lexical database. Based on a contemporary printed corpus of over 227 million words, P-PAL provides a broad range of word attributes and statistics, including several measures of word frequency (e.g., raw counts, per-million word frequency, logarithmic Zipf scale), morpho-syntactic information (e.g., parts of speech [PoSs], grammatical gender and number, dominant PoS, and frequency and relative frequency of the dominant PoS), as well as several lexical and sublexical orthographic (e.g., number of letters; consonant–vowel orthographic structure; density and frequency of orthographic neighbors; orthographic Levenshtein distance; orthographic uniqueness point; orthographic syllabification; and trigram, bigram, and letter type and token frequencies), and phonological measures (e.g., pronunciation, number of phonemes, stress, density and frequency of phonological neighbors, transposed and phonographic neighbors, syllabification, and biphone and phone type and token frequencies) for ~53,000 lemmatized and ~208,000 nonlemmatized EP word forms. To obtain these metrics, researchers can choose between two word queries in the application: (i) analyze words previously selected for specific attributes and/or lexical and sublexical characteristics, or (ii) generate word lists that meet word requirements defined by the user in the menu of analyses. For the measures it provides and the flexibility it allows, P-PAL will be a key resource to support research in all cognitive areas that use EP verbal stimuli. P-PAL is freely available at http://p-pal.di.uminho.pt/tools.

Keywords Lexical databases · Word frequency · Orthographic word statistics · Phonological word statistics · European Portuguese

Advances in psycholinguistics have been accompanied by an increasing demand for the control of word properties, which has been made possible through the development of lexical databases that provide researchers with information about the structural (attributes) and distributional (statistics) characteristics of words in a given language. The first attempts to develop these databases date back to 1921 when Thorndike published *The Teacher's Word Book*, a work that ranked the most frequent 10,000 English words on the basis of the manual count of the number of times a given word occurred in an English corpus of 4.5 million words. Since then, the dramatic advances in technology have it made possible to collect larger and larger amounts of words from an increasing number of linguistic sources and registers. These included written texts from literature, textbooks, technical reports, newspapers, transcriptions of spoken productions, and, more recently, from film and television subtitles, which have proved to be a relevant determinant of the speed and accuracy with which words are named and/or recognized in different languages (see Soares et al., 2015, for a recent review).

Lexical databases also began to offer an increasing number of word statistics. Indeed, besides the computation of the number of times a given word appears in a language (i.e., its frequency of use) as in Thorndike's seminal work, lexical databases started

✉ Ana Paula Soares
asoares@psi.uminho.pt

[1] Human Cognition Lab, CIPsi, School of Psychology, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

[2] Centre for Humanistic Studies, University of Minho, Braga, Portugal

[3] Computer Science and Technology Center, University of Minho, Braga, Portugal

[4] Computational Linguistics Department, University of Tübingen, Tübingen, Germany

[5] Department of Spanish & Portuguese, University of Toronto, Toronto, Ontario, Canada

[6] Department of Methodology, University of València, Valencia, Spain

Published online: 31 May 2018

to provide other word attributes such as word length (in number of letters, phonemes and syllables), pronunciation, stress pattern, part-of-speech [PoS] information, and also other measures aiming at capturing the degree of similarity (orthographic and/or phonological) among words in the lexicon (for instance, the word "fall" is visually similar to words such as "call," "mall" and "fell," as the word "gate" sounds like "hate" and "bait") in the so-called neighborhood statistics as the classic *N* metric of Coltheart, Davelaar, Jonasson, and Besner (1977; see also Luce & Pisoni, 1998, for an equivalent measure in the phonological domain), and, more recently, the orthographic Levenshtein distance (OLD20) proposed by Yarkoni, Balota, and Yap (2008). Moreover, refined measures of word frequency such as the number of different contexts in which a word appears (e.g., Adelman, Brown, & Quesada, 2006; see also Perea, Soares, & Comesaña, 2013, and Parmentier, Comesaña, & Soares, 2017), the logarithmic Zipf scale measure (e.g., van Heuven, Mandera, Keuleers, & Brysbaert, 2014), or the distribution of word frequencies according to the PoS information (e.g., Baayen, Piepenbrock, & van Rijn, 1993; Balota et al., 2007; Brysbaert, New, & Keuleers, 2012; Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013; Kyparissiadis, van Heuven, Pitchford, & Ledgeway, 2017; New, Pallier, Brysbaert, & Ferrand, 2004) were also made available in lexical databases (the word "play," for example, can appear in a corpus as a noun or as verb, each with a different number of occurrences in the English language).

At a sublexical level, databases also offer a broad range of statistics targeting word subcomponents such as the number and the frequency of syllables, morphemes, bigrams (co-occurrences of two letters), letters, biphones (co-occurrences of two phones), phones, or the probability with which different phonological or orthographic segments (letters/phones, bigrams/biphones, syllables) occur in a given language (e.g., Baayen, Feldman, & Schreuder, 2006; Balota et al., 2007; Bédard et al., 2017; Boudelaa & Marslen-Wilson, 2010; Chetail & Mathey, 2010; Davis, 2005; Davis & Perea, 2005; Duchon et al., 2013; Duñabeitia, Cholin, Corral, Perea, & Carreiras, 2010; Hofmann, Stenneken, Conrad, & Jacobs, 2007; Ktori, van Heuven, & Pitchford, 2008; Kyparissiadis et al., 2017; New & Spinelli, 2013).

The control and/or manipulation of all these word attributes and statistics assume a major role in research, since studies conducted in the last decades have shown that they affect word processing (see Balota, Yap, & Cortese, 2006, and Yap & Balota, 2015, for reviews), although the magnitude and the direction of the effects seem to depend on the specificities of each language (e.g., it is well known that the regularity of the spelling-to-sound correspondences affects the type of effects that can be observed across languages, with larger frequency and lexicality effects observed in languages with more opaque writing systems, and stronger phonological effects in languages with more shallow orthographies; see Frost, Katz, & Bentin, 1987; Goswami, Ziegler, Dalton, & Schneider, 2001; Grainger & Ziegler, 2011). Therefore, and despite the differences observed in the effects of these variables across languages, an issue that is beyond the scope of this article, what we intend to emphasize here is that accumulated evidence clearly demonstrates that words are extremely complex stimuli and acknowledging it is critical for conducting well-controlled and well-designed research not only in psycholinguistics, but in all the research areas that use verbal stimuli. Hence, developing lexical databases that provide reliable information about word attributes and statistics in a given language is not only a desirable goal, but a key requirement for current research.

However, although these databases are available for languages like English (e.g., MRC: Coltheart, 1981; CELEX: Baayen et al., 1993; N-Watch: Davis, 2005), French (e.g., LEXIQUE: New et al., 2004; InfoSyll: Chetail & Mathey, 2010; Diphones-fr: New & Spinelli, 2013; SyllabO+: Bédard et al., 2017), Dutch and German (e.g., CELEX: Baayen et al., 1993; DlexDB: Heister et al., 2011), Spanish (e.g., LEXESP: Sebastián-Gallés, Martí, Cuetos, & Carreiras, 2000; BuscaPalabras: Davis & Perea, 2005; EsPal: Duchon et al., 2013; SYLLABARIUM: Duñabeitia et al., 2010), Greek (e.g., GreekLex: Ktori et al., 2008; Kyparissiadis et al., 2017), Basque (E-Hitz: Perea et al., 2006) or Arabic (e.g., ARALEX: Boudelaa & Marslen-Wilson, 2010), they are scarce for European Portuguese (EP). Until 2000, the only lexical database available for EP was the Português Fundamental [Fundamental Portuguese, FP] (Nascimento, Marques, & Cruz, 1987; Nascimento, Rivenc, & Cruz, 1987), a work providing frequency measures for 2,217 EP words drawn from a small EP spoken corpus (700,000 words) compiled during the 1970s. In 2000, Nascimento, Pereira and Saramago developed the Léxico Multifuncional Computorizado do Português Contemporâneo [Multifunctional Computational Lexicon of Contemporary Portuguese, MCL], providing frequency norms for 26,443 lemmatized and 140,315 nonlemmatized EP word forms extracted from a larger (~16 million) EP printed corpus named Corlex (see Nascimento, Pereira, & Saramago, 2000, for details). Lemma databases (i.e., databases offering word statistics based on the canonical form of words; e.g., the lemma "play" represents the inflected forms "play," "plays," "played," "playing") have become popular since Baayen, Dijkstra, and Schreuder (1997) showed that lemma counts were more informative than word form counts (i.e., a word as it appears in its "natural" form; e.g., occurrences of "play," "plays," "played," or "playing" separately) in word recognition. Note, however, that since lemma counts were based on the summed frequencies of all the inflected forms integrated in the same lemma, they tend, on the one hand, to overestimate the number of times a given word appears in its "natural" form in a corpus (particularly in highly inflected languages such as EP), and, on the other hand, to

underestimate the frequency with which sublexical units (e.g., bigrams) occur in the inflected forms (see Hofmann, Stenneken, Conrad, & Jacobs, 2007, for similar arguments). For example, in Procura-PALavras (P-PAL), the frequency of the lemma *jogar* [play] corresponds to 199,6379 occurrences per million words (pmw), whereas the word form frequency of *jogar* corresponds to 94,2041 occurrences pmw. Conversely, the summed frequency of the bigram "jo" corresponds to 230,303 pmw in the P-PAL lemma database and to 379,367 pmw in the P-PAL word form database. Thus, due to these biases, word form databases are increasingly recommended. Furthermore, subsequent studies have also shown that, contrary to Baayen et al.'s (1997) findings, word form frequencies account for slightly more variance in word recognition times than do lemma frequencies (e.g., Brysbaert & New, 2009; Brysbaert et al., 2012). Nevertheless, in P-PAL, as in other lexical databases (e.g., CELEX, Lexique, GreekLex, E-Hitz), lemma and word form measures are provided, hence leaving researchers free to choose which database they want to use in their studies.

In addition, it is also worth noting that, regardless of the use of lemma or word form counts, another important issue concerns the dimension of the corpus from which word counts were extracted. To obtain reliable word estimates, recent studies suggest using corpora of at least 20–30 million words (see for instance Brysbaert & New, 2009, or Brysbaert et al., 2011). Computing word frequency from a smaller corpus, tends to underestimate the number of times words occur in a language, especially low frequency words. This aspect is particularly critical, since recent reports have shown that almost the entire word frequency effect in word recognition lies in words below ten occurrences pmw ($\log_{10} = 1$), with the most significant effect being observed for words with a frequency between 0.1 ($\log_{10} = -1$) and 1 ($\log_{10} = 0$) pmw (see Balota et al., 2007, and also Brysbaert et al., 2011, for details). Thus, even though the EP word counts provided by the MCL database (Nascimento et al., 2000) were based on larger corpora than the EP word counts provided by the FP database (Nascimento, Marques, & Cruz, 1987; Nascimento, Rivenc, & Cruz, 1987), both are below the recommended dimensions. Moreover, besides word frequency, these databases only provide PoS information for each of their lexical entries, which is a significant obstacle for the conduction of research with EP verbal stimuli, as they do not allow for the proper control of all other lexical and/or sublexical variables affecting word processing (e.g., Balota et al., 2006; Yap & Balota, 2015).

Acknowledging these limitations, Gomes and Castro (2003) developed Porlex, an EP database offering orthographic, phonological, phonetic, PoS, and neighborhood statistics for ~30,000 words (uninflected content words and inflected function words). However, despite its relevance, Porlex provides word frequency for only 5% of its lexical entries (~1,500 words) obtained from the FP database (Nascimento, Marques, & Cruz, 1987; Nascimento, Rivenc, & Cruz, 1987), which, as mentioned, is an outdated and very small EP spoken corpus (less than 1 million words). For this reason, all the Porlex lexical and sublexical statistics contained a serious bias.

Nowadays several EP resources provide word frequency measures from large-scale corpora. For instance, from the Linguateca resource center (see www.linguateca.pt/), it is possible to obtain EP raw word counts from 19 different corpora varying in literacy genre and register (e.g., the Vercial corpus contains records from EP archaic texts indexed from the 16th to the 20th century, or the Museu Pessoa corpus, which contains spoken records from interview transcriptions conducted both in Portugal and Brazil). However, in this online resource center, only two word queries are possible, namely either searching for word frequency in a specific corpus or in all corpora at once. This inevitably results in an EP word frequency measure that contains incidences from archaic EP and from Brazilian Portuguese, or in an EP word frequency measure that is exceedingly dependent on the type of language register from which the word counts were obtained. Indeed, since word frequency aims to capture the "real" use that native speakers make of their language, it is critical that the corpus from which word counts are drawn is not only large enough, but, importantly, as varied as possible in its internal composition (see Sinclair, 2005, or Brysbaert et al., 2011). Register diversity would increase language representativeness and, hence, the number of reliable lexical and sublexical measures (see Soares et al., 2014).

Bearing these issues in mind, we developed P-PAL, a four-year research project that aimed to offer the scientific community a Web-based application with a broad range of frequency, morpho-syntactic, orthographic and phonological word attributes and statistics with different grain sizes (word as a whole, syllables, trigrams, bigrams, letters, biphones, and phones) not yet available for EP, and obtained from a large-size (over 227 million words) and diversified (including records from spoken and written texts from diverse genres) contemporary EP corpus. It is the outcome of this project that we present in this article. We begin by describing corpus sampling procedures and by characterizing the indexation of the lexical entries in lemma and word form databases. Then, we present the Web-based interface developed, as well as the word attributes and statistics provided.

Corpus sampling

For the creation of the P-PAL corpus and for the computation of all lexical and sublexical measures provided in its Web-based interface, eight morpho-syntactically tagged EP corpora were compiled: seven from the Linguateca language resource center (CETEMPúblico, DiaCLAV, Avante!, Natura/Minho,

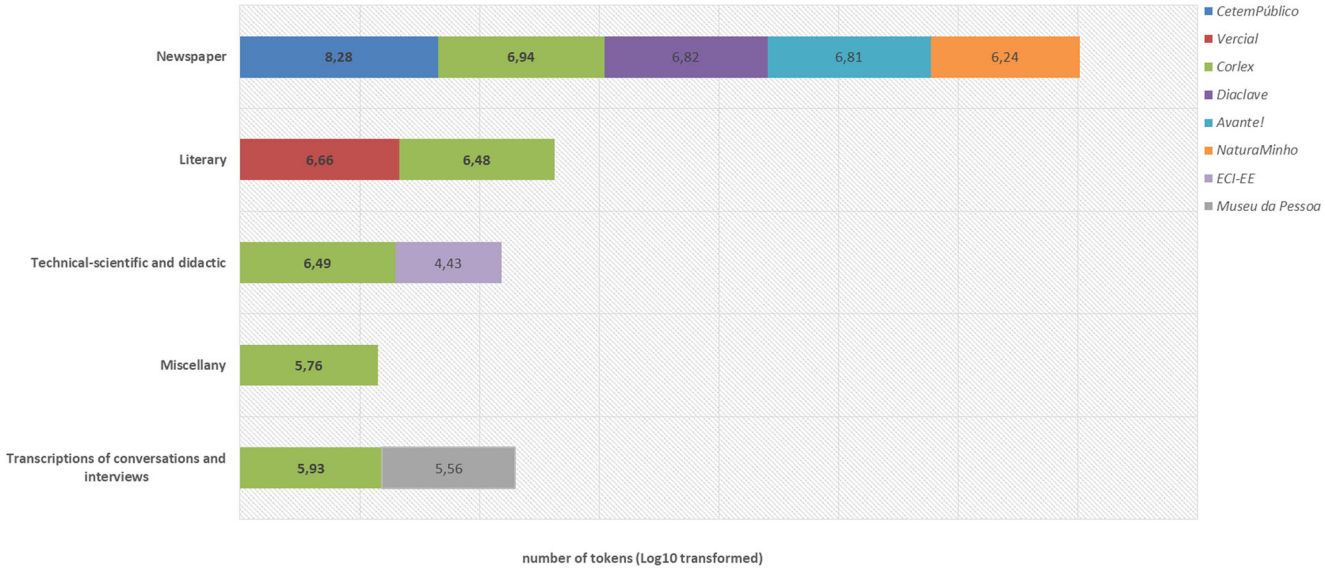


Fig. 1 Type and genre distribution of the P-PAL corpora. Note that all numbers within the bars are $\log_{10}$-transformed.

ECI-EE, Museu da Pessoa, and Vercial),[1] along with the Corlex corpus, from which the MCL word counts (Nascimento et al., 2000) were drawn (see Soares et al., 2014, for a detailed description of each P-PAL subcorpus). The compilation of all these corpora resulted in a megacorpus of 227,770,752 occurrences (tokens), 226,552,040 of which were from EP written texts of different genres (e.g., newspapers, literary, technical–scientific, and didactic texts), and the remaining 1,218,712 were from orthographic transcriptions of informal conversations and more formal spoken productions, such as at conferences or in radio and television interviews. Figure 1 presents the distribution ($\log_{10}$ transformed) of the types of language registers and genres in the P-PAL subcorpora.

As can be observed in Fig. 1, most of the P-PAL corpus consists of written records from newspapers (94.5% of the total corpus). In this genre, CETEMPúblico contributes with the most significant number of occurrences (89.1%), followed by Corlex (4%), DiaCLAVE (3.1%), Avante! (3%), and Natura/Minho (0.8%). The literary genre represents 3.4% of the total corpus, the highest proportion of occurrences (60%) stemming from Vercial. The technical–scientific and didactic genres represent 1.6% of the total corpus, with Corlex contributing with the most significant portion (99.3%). The ECI-EE accounts for only 0.7% of occurrences. The "miscellaneous" genre from Corlex includes 575,962 occurrences, corresponding to 0.3% of the total written corpus. Although in the P-PAL corpus the distribution of the different registers and genres is not balanced (with the vast majority of registers coming from newspaper texts), the inclusion of several newspapers from different regions in Portugal (from north to south, including the islands) covering a wide variety of themes (e.g., Avante! is a newspaper corpus that collects texts of political content; see Soares et al., 2014, for details) was intentionally done in the P-PAL corpus to best represent the diversity of the EP language. One might argue that it would have been desirable to include records from a wider variety of sources (e.g., literary texts, legal texts, academic texts, among others). Here, we focused on creating a lexical database for contemporary EP, gathering uncopyrighted records, whose content resembles the everyday use of language as closely as possible. Note that literary texts, legal texts or academic reports usually resort to uncommon and often outdated words, which contribute both to overestimate the number of times rare words appear in the corpus, and to underestimate the number of times more common words appear in the same corpus, hence introducing a bias in the frequency measures obtained from these corpora (see Baayen, 2011; Breland, 1996; Brysbaert et al., 2011; Soares et al., 2014; Soares et al., 2015). Furthermore, it is not uncommon for other databases to include asymmetrical genre types (e.g., in the EsPal database (Duchon et al., 2013), around 44% of the corpus was gathered from Wikipedia), and although most of the P-PAL corpus comprises newspaper records, Soares et al. (2015) showed that the P-PAL word frequency accounts for percentages of variance similar to those observed in other international written-text databases (e.g., CELEX, British National Corpus, Lexique 2, and EsPal; see Brysbaert & New, 2009; Duchon et al., 2013; Keuleers, Brysbaert, & New, 2010; New, Brysbaert, Veronis, & Pallier, 2007; van Heuven et al., 2014). Hence, we believe this choice has not affected the characteristics of the final P-PAL database—namely the types of words included—as can be inferred from the frequency distributions presented in Figs. 2 and 3.

[1] Note that although the Museu Pessoa corpus includes, in its original version, spoken records from Brazilian Portuguese, in the P-PAL corpus we have only considered spoken records from the European Portuguese (EP) variant.

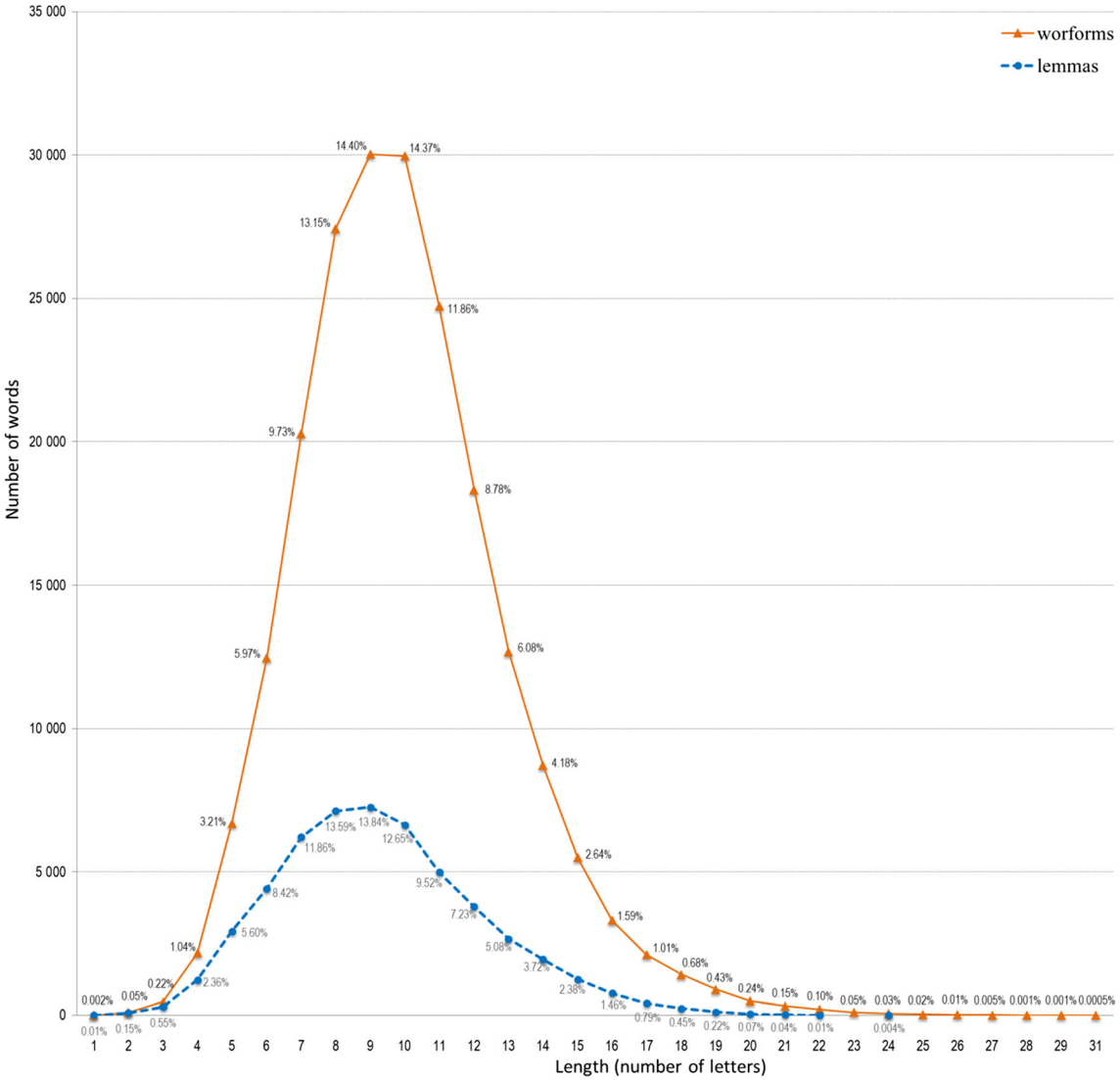


Fig. 2 Distribution of word length for the 208,642 word forms and 52,404 lemmas in P-PAL (numbers of words as a percentage of all the words of each length in the database are also presented).

Lemma and word form databases

Before computing the frequency, morpho-syntactic, orthographic, and phonological statistics provided in the application, the lexicon from the lemma and word form corpora was indexed. Since each of the eight subcorpora integrated in P-PAL were already morpho-syntactically tagged and lemmatized, we began by developing a morpho-syntactic system to accommodate the grammatical classifications adopted in each, in accordance with the PoS categorization of Casteleiro (2001; for details about the indexation procedures adopted see Soares et al., 2014). Thus, in P-PAL, lemmas and word forms were automatically assigned the following main PoS categories: nouns (N), adjectives (ADJ), verbs (V), adverbs (ADV), conjunctions (CONJ), determiners (DET), interjections (INT), quantifiers (QUANT), prepositions (PREP), and pronouns (PRON). Moreover, DET, PRON, QUANT ADV, and CONJ were additionally classified into PoS subcategories. Specifically, DETs were subclassified as demonstrative (DET_dem), possessive (DET_poss), indefinite (DET_ind), relative (DET_rel), interrogative (DET_inter) and the articles as definite (Art_def), or indefinite (Art_ind); PRONs as personal (PRON_pers), demonstrative (PRON_dem), indefinite (PRON_ind), possessive (PRON_poss), interrogative (PRON_inter), and relative (PRON_rel); QUANTs as universal (QUANT_univ), existential (QUANT_exist), relative (QUANT_rel), interrogative (QUANT_inter), cardinal number (Num_card), ordinal number (Num-ord), fractional number (Num_frac), and multiplicative number (Num_mult). Finally, ADVs were subclassified as interrogative (ADV_inter) and CONJ as subordinating (CONJ_sub) and coordinating (CONJ_coord). After cross-checking the grammatical information with the JSpell automatic analyzer (Simões & Almeida, 2001) and

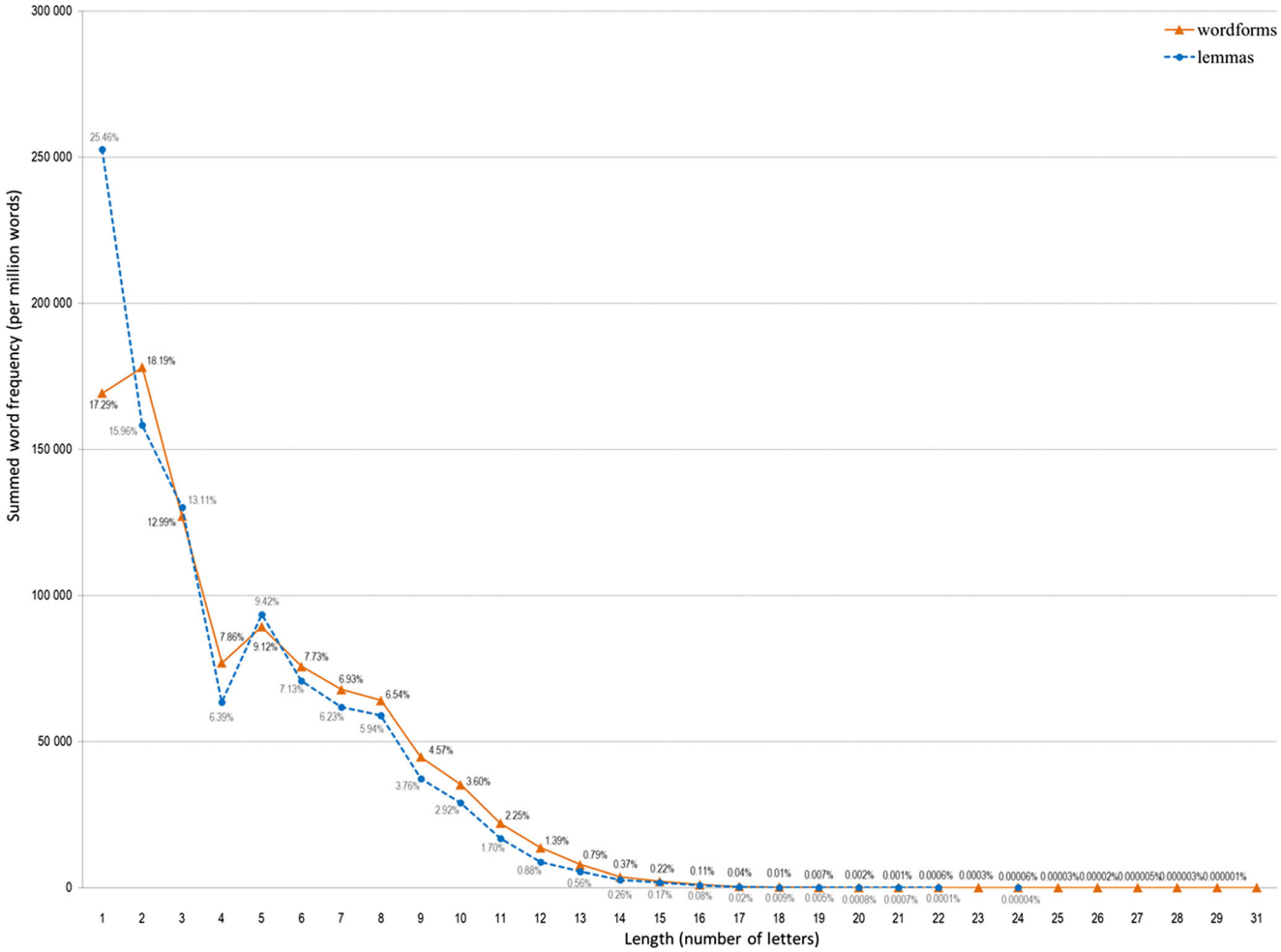


Fig. 3 Summed word frequencies (per million occurrences) for the 160,604 word forms and 41,500 lemmas in P-PAL, as a function of word length (word summed frequencies as a percentage of all the words of each length in the database are also presented).

manually verifying the tags whose PoS information was inconsistent, the final indexation of the P-PAL lexicon (lemma and word form) was conducted. This resulted in 52,404 different lexical entries in the lemma database and 208,642 different lexical entries in the word form database, as is displayed in Fig. 2. Note that in the P-PAL lemma database, the infinitive form of the verb (e.g., *ser* [to be]) was chosen to represent all inflected forms of the verbal paradigm (e.g., *sou* [I am], *és* [you are], *é* [is], and *era* [was]). The categories N and ADJ are represented by the masculine singular form (e.g., *menino* [boy], *bonito* [pretty]), which comprises the entire nominal (e.g., *menino* [boy], *menina* [girl], *meninos* [boys], *meninas* [girls]) or adjectival (e.g., *bonito* [pretty], masculine, singular; *bonita* [pretty], feminine, singular; *bonitos* [pretty], masculine, plural; *bonitas* [pretty], feminine, plural) paradigm. In strictly masculine or feminine nouns, the singular form is used (e.g., *animal* [animal], *comboio* [train], *costa* [coast], *adivinha* [riddle]). Singular feminine words with different stems have also been included as different lemmas (e.g., *homem* [man], *mulher* [woman]). In the word form P-PAL database, all the different inflected forms of given words were indexed.

As is illustrated in Fig. 2, P-PAL includes, in the lemma database, words ranging in length from 1 to 24 letters, and in the word form database, words ranging from 1 to 31 letters. Most of the words in both databases are between 7 and 11 letters long, which represents 61.5% and 63.5% of the entire lexicon, respectively. The mean numbers of letters are 9.3 letters (*SD* = 2.96) in the lemma database and 9.9 letters (*SD* = 2.97) in the word form database. This high number of letters per word reflects the fact that EP is a agglutinate language, in which words can be created not only by adding prefixes and/or suffixes, but also by compounding two or more morphemes (including stems and affixes) into one single word while maintaining the original morphemes relatively unchanged. For instance, the 24-letter lemma *socialista–revolucionário* that rarely occurs in EP (0.0141 occurrences pmw) or the 31-letter word form *integracionistas–centralizadoras* that occurs at a frequency of 0.0049 pmw, resulted, in both cases, from the junction of two distinct compound EP words: *socialista* is a

compound word that is formed by the word *social* [social] and the suffix *-ista*, which denotes the adoption of a doctrine, theory, or political system, with the word *revolucionário* that is also an EP compound word formed by the word *revolução* [revolution] and the suffix *-ário*, which denotes someone who performs a particular action (an agent). The same is observed for the word form *integracionistas–centralizadoras*, which entails the EP compound word *integracionista*, formed by the junction of the word *integrar* [to integrate], plus the suffixes *-ção* (denoting a state of being) and *-ista* (denoting the adoption of a doctrine or a belief, as we mentioned above), with the EP compound word *centralizadoras*, formed by combining the word *central* [central] plus the suffixes *-izar* (denoting to become) and *-dor(as)* denoting a state or a quality. Figure 3 shows the distributions of the summed word frequencies (pmw) in the P-PAL lemma and word form databases as a function of word length (number of letters).

The distribution of the summed frequencies in P-PAL reveals a Poisson-like distribution, in both the lemma and word form databases, as has been observed in other languages (see Grzybek, 2006, for a review). This distribution reveals that as the number of letters in the P-PAL words (lemmas or word forms) increases, the probability of occurrence of the word decreases. Furthermore, the distribution analysis reveals that more than 50% of the lexical frequencies in the lemma database occur for words with three or fewer letters (54.53%), with about 90% of occurrences observed for words up to nine letters. In the word form database, a similar distribution was observed with 56.32% of frequencies observed for words with four or fewer letters and 93.40% of occurrences also for words up to nine letters. One-letter words present the highest summed word frequency in the lemma database, since they include the three functional EP words *a* (PREP meaning "to"; Art_def; PRON, feminine of "the"), *o* (Art_def, PRON, masculine of "the"), and *e* (CONJ_coord, meaning "and"), with pmw frequencies of 88,046.59, 80,466.16, and 84,061.31, respectively. In the word form database, two-letter words comprise the set of the most frequent words, among which the functional words *de* (PREP, meaning "of") and *em* (PREP, meaning "in"/"on"/"at") are included, with pmw frequencies of 46,474.75 and 12,561.91, respectively, followed closely by one-letter word forms, which include the functional words *a*, *o*, and *e*, as in the lemma database, plus the word forms *à* (PREP *a* + Art_def *a* or PRON *a*, meaning "to the") and *é* (third person singular of the verb *ser* [to be], meaning "is"), with pmw frequencies of 39,164.26, 30,020.17, 87,551.52, 5,050.34, and 7,391.74, respectively.

Web-based interface

The P-PAL Web-based interface was designed to be a user-friendly application to allow researchers from all areas of study that use EP verbal materials (e.g., psycholinguistics, linguistics, memory, neurosciences) to access a broad range of word attributes and lexical and sublexical statistics not yet available for EP in a quick and efficient way. The P-PAL interface is freely available for research purposes at http://p-pal.di.uminho.pt/tools.

When the user enters the application, a dialog box appears asking the user to specify which of the two word queries available he/she wants to perform: (i) to analyze words previously selected by the researcher in specific attributes and lexical and/or sublexical characteristics, or (ii) to generate word lists that meet specific word requirements defined by the user in the menu of analysis. Then, regardless of the word query selected, users should decide in which of the P-PAL databases (i.e., lemma or word form) they want to conduct their word search, as is illustrated in Fig. 4.

After the user decides on the word query and database, the analysis menu is displayed (see Fig. 5). This brings up a list of all the attributes and statistics available in the interface, irrespective of the word query and database chosen. Nevertheless, if the user chooses to conduct an "analyze word list" query, he or she will be additionally required to upload a file (i.e., a text file [.txt] or an Excel file [.xls] with the ISO-8859-1 or UTF-8 file encoding) containing the words to be analyzed by the application in the attributes/statistics selected. If a "generate word list" query is chosen instead, users will be required to define the attributes/statistics that the words should meet. For instance, if the user intends to obtain words whose lexical frequency ranges between 1 and 10 pmw, he or she should specify this in the constraints field associated with the pmw frequency measure by typing "1" in the minimum (Min.) and "10" in the maximum (Max.) values of the interval associated with that measure (the maximum, minimum and mean values obtained for each statistic are also provided to guide word constraints; see Fig. 5). Thus, the same interface is provided

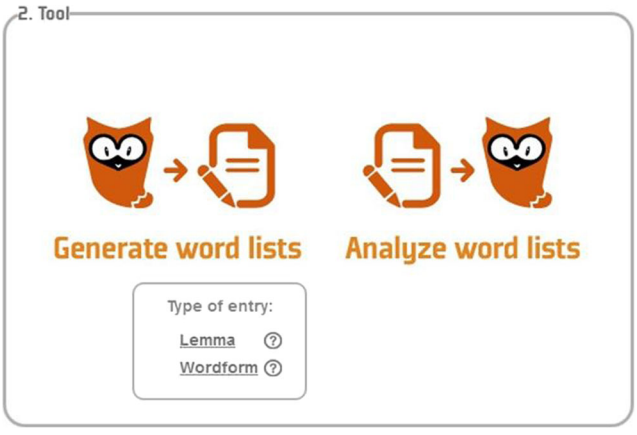


Fig. 4 Depiction of the word query menu. On the left, the lemma and word form queries apply to the "generate word lists" option, and on the right the same options (not visible in the figure) are available for the "analyze word lists" option.

Fig. 5 Depiction of the word frequency measures available in the P-PAL Web-based interface. Note that this illustrates a "generate word list" query in the application, but the same statistics can be observed for the "analyze word list" query, except that the constraint options are not presented.

whether the user is conducting an "analyze word list" or a "generate word list" word query, with the exception that only in the latter are the constraint fields displayed. Note that all metrics in the "analyze word list" query are also available in the "generate word list" functionality, but in the latter, the user is provided with constraints fields, in which value ranges or specific terms can be inserted/selected. However, constraints are not available for certain PoS and neighborhood metrics—namely, those regarding the secondary (remaining) grammatical categories of a word, the percentage of occurrence of these secondary categories in the corpus, and the list of orthographic and phonological neighbors (addition, deletion, transposed and phonographic neighbors) in the corpus (see ahead for a detailed description of each of these word attributes and statistics).

Note that although analyzing word lists is a word query available in the majority of international lexical databases that have been developed so far (e.g., N-Watch, CELEX, Lexique, BuscaPalabras), getting words that meet specific requirements is an option hardly found (e.g., in the English Lexicon Project or EsPal), but this is strongly recommended. Indeed, creating such constraints will contribute not only to optimizing the stimulus selection—allowing, for instance, researchers to save a considerable amount of time searching and matching stimuli for several lexical and sublexical characteristics—but also to minimizing errors in that process. Moreover, this functionality might also contribute to reducing the experimenter bias often observed when researchers assume the responsibility for selecting on their own the experimental items to be used in a given experiment, even without consciousness or intention of doing so (see, e.g., Forster, 2000). Therefore, providing a "generate word list" option and combining it with the traditional "analyze words lists" query in a single application, such as in P-PAL, is an important feature that gives strong versatility to a research tool and increases its usefulness in supporting research with EP verbal stimuli.

Word attributes and statistics

When the analysis menu is displayed, the only statistic selected by default is the pmw frequency (see Fig. 5), due to the importance of this variable in all studies using verbal stimuli (see Brysbaert et al., 2011, or Soares et al., 2015, for recent reviews). All the other word attributes/statistics in which the user is interested should be selected by selecting the checkbox to the left of each word property in the analysis menu.

In P-PAL, word attributes and statistics are organized into four main fields that entail several lexical and sublexical measures of different grain sizes (word as a whole, syllables, trigrams, bigrams/biphones, letters/phones)—namely, (i) word frequency measures, (ii) morpho-syntactic information, (iii) orthographic statistics, and (iv) phonological statistics (see Fig. 5 for an illustration). The option to organize the word attributes/statistics into these four broad fields relies on the fact that the majority of researchers working with verbal stimuli are interested in obtaining word properties/statistics based on either their visual (orthographic) or their spoken (phonological) forms, thus making it easier to search for these attributes/statistics in the application. So, researchers interested in studying written language processing or processes that depend mainly on words' visual features are strongly encouraged to collect word attributes/statistics from the orthographic field. Conversely, researchers interested in studying spoken language processing or processes that depend mainly on the phonological properties of EP words are encouraged to obtain

these attributes/statistics from the phonological field. Some researchers might also be interested in collecting metrics from both the orthographic and the phonological fields, due to the accumulated evidence suggesting, for example, that phonological codes are activated during the visual recognition of printed words (e.g., Goswami et al., 2001; Grainger & Ziegler, 2011). Since word frequency statistics and word PoS information interest, in principle, researchers from all the areas of inquiry, these metrics are provided separately in the word frequency and in the morpho-syntactic fields, respectively. The word attributes and statistics (lexical and sublexical) included in each of these fields are described below.

Word frequency measures Seven word frequency measures are available in P-PAL (see Fig. 5). In addition to the classic pmw frequency measure (freq_corp_mil) previously mentioned (in the lemma database, freq_corp_mil ranges from 0.0047 to 89,567.7033, M = 19.0224, and in the word form database it ranges from 0.0049 to 87,551.5215, M = 4.8273), P-PAL also provides the raw measure of the number of times a given word occurs in the lemma or word form corpus (freq_corp_abs). In the word form database, the raw frequency values range from a minimum of 1 to a maximum of 17,921,249 occurrences (M = 988.13 occurrences), whereas in the lemma database they range from 1 to 19,084,706 (M = 4,053.2191). Note that the word frequency statistics provided in this field correspond to the number of times a given word occurs in the corpus (lemma or word form), irrespective of its syntactic role in each case—that is, its PoS categorization. Thus, the lexical frequency of the word form "play," for instance, results from the summed frequencies of "play" as both a verb and a noun in the word form corpus. Similarly, the lemma frequency of "play" results from the sum of all the inflections that "play" presents, both as a verb (e.g., play, plays, played) and as a noun (e.g., play). Nevertheless, it is worth noting that although the word frequencies provided in this field correspond to the sum of the frequencies of the same lemmas/word forms, regardless of PoS information, it is possible to obtain word frequencies disambiguated by PoS category in the morpho-syntactic field, as we describe below (i.e., to obtain separately word frequencies for the lemma or word form of "play" as a verb and as a noun). Additionally, measures as the $\log_{10}$ of the number of times a word appears in the lemma or word form corpus (freq_corp_abs_log) and the $\log_{10}$ of the pmw frequency after summing 1 to the pmw frequency value (freq_corp_log) are also provided (in the lemma database freq_corp_abs_log ranges from 0.3010 to 7.2807, M = 1.7209, and freq_corp_log ranges from 0.0020 to 4.9522, M = 0.3087, whereas in the word form database freq_corp_abs_log ranges from 0.3010 to 7.2534, M = 1.1978, and freq_corp_log from 0.0021 to 4.9423, M = 0.1445). Adding 1 to the number of occurrences (the Laplace transformation) precludes the existence of negative values for low-frequency words. Furthermore, doing so makes it possible to match stimuli from different corpora when a stimulus is not present in any of them, as was suggested by Brysbaert and Diependaele (2013). In addition, the $\log_{10}$ of the pmw frequency + 1 divided by the number of words in the corpus expressed in millions (freq_corp_mil_log_n) is also provided, to correct for differences in corpus size, as was suggested by Brysbaert et al. (2011). Also, the squared $\log_{10}$ of the pmw frequency + 1 (freq_corp_mil_log_sqr) is provided, because the relationship between log frequency and word latencies is not completely linear and is captured better by the log square value, as several studies have demonstrated (see, e.g., Baayen et al., 2006; Brysbaert & New, 2009; Soares et al., 2015); in the lemma database, freq_corp_mil_log_n ranges from – 2.0271 to 4.9522, M = – 0.6075, and freq_corp_mil_log_sqr ranges from 0 to 24.5243, M = 0.3533, whereas in the word form database, freq_corp_mil_log_n ranges from – 2.0097 to – 4.9423, M = – 1.1131, and freq_corp_mil_log_sqr ranges from 0 to 24.4263, M = 0.1288.

Finally, P-PAL also offers the standardized Zipf scale measure (freq_corp_zipf) for each of its lexical entries (lemmas and word forms), calculated by adding 3 to the $\log_{10}$ of the per-million-word frequency (see van Heuven et al., 2014, for details). The Zipf scale is assumed to be a much easier and more intuitive way to understand the word frequency distribution, since it depicts word frequencies on a logarithmic scale, similar to the decibel scale. In the P-PAL lemma database, freq_corp_zipf ranges from 0.6721 to 7.9522, M = 2.3314, and in the P-PAL word form database it ranges from 0.6602 to 7.9423, M = 1.7771. Since content words (e.g., ADJ, N, V) typically present Zipf values lower than 6 (note that in both databases all words with a Zipf value above 6 correspond mainly to function words, such as the words *o*, *a*, *e*, *de*, or *em* previously mentioned), for the majority of research purposes the Zipf scale ranges between 1 and 6. Words presenting a Zipf value from 1 to 3 are considered low-frequency words (with frequencies of 1 per million words or below), whereas words with a Zipf value above 4 are considered high-frequency words (with frequencies of 10 per million words or higher). Note, however, that words presenting a Zipf value below 1 (corresponding to 1.05% of the lexical entries in the lemma database and to 2.18% of the lexical entries in the word form database) are rarely used in the language and presumably would be unknown to the majority of native EP speakers.

Morpho-syntactic information P-PAL presents ten PoS measures, shown in Fig. 6. Specifically, in this field, P-PAL offers information regarding the main PoS category (morf_cat) and/or the subcategory (morf_type) that each of its lexical entries assumes in the lemma or word form corpus (bear in mind that the words are classified according to the categories of N, ADJ,

Fig. 6 Depiction of the morpho-syntactic measures available in the P-PAL Web-based interface. Note that this illustrates a "generate word list" query in the application, but the same statistics can be observed for the "analyze word list" query, except that the constraint options are not presented.

V, ADV, CONJ, DET, INT, QUANT, PREP, and PRON, and, additionally, that DET, PRON, QUANT, ADV, and CONJ were further subclassified into other grammatical subclasses; see the Corpus Sampling section above), in line with recent databases that provide PoS information (e.g., Brysbaert et al., 2012; Duchon et al., 2013; Kyparissiadis, et al., 2017). Because syntactic ambiguity is very common in EP (e.g., the word *crítico* [critic, critical] can be used both as a noun and an ADJ), P-PAL displays all the grammatical categories that have been assigned to each lexical entry according to their frequency of occurrence (descending order). PoS tags are comma-separated. Moreover, information concerning the most frequent category observed in the corpus (morf_max_cat), the percentage with which the higher-frequency grammatical category occurs (morf_max_d), the frequency (pmw) of the higher-frequency grammatical category (morf_max_freq), and information regarding the remaining PoS categories (morf_others_cat) and their relative distribution (%) in the corpus (morf_others_d) are also provided. As an example of the PoS information provided in P-PAL, the output information for the word form *crítico* is the following: morf_cat = ADJ, N, indicating that *crítico* occurred in the ADJ and N categories in the word form corpus; morf_type = NONE, since ADJ and N have no grammatical subclasses; morf_max_cat = ADJ, indicating that the most frequent grammatical category of *crítico* is ADJ; morf_max_d = 72.69, showing that *crítico* as an ADJ occurs in ~73% of the occurrences in the word form corpus; morf_max_freq = 24.2509, indicating the pmw frequency of *crítico* as an ADJ; morf_others_cat = N, showing that besides occurring as an ADJ, *crítico* also occurs as an N in the word form corpus; and morf_others_d = 27.31, indicating that *crítico* as an N occurs ~27% of the time. Similarly, in the lemma database, the user can access all the grammatical categories assigned to a given lemma. For instance, although the frequency of the lemma *crítico* is 104.7139 pmw, it is possible to observe that *crítico* occurs both as an ADJ and an N, and that the pmw frequency of the most frequent PoS (ADJ) is 79.7134 pmw occurring ~76% of the time in the lemma corpus. Information concerning the distribution of the other PoS categories (N) is also provided, as in the word form example presented. Hence, even though the frequency counts obtained from the word frequency field in the lemma and word form databases combine all frequency values regardless of their grammatical class, in this field, users can access word frequency statistics disambiguated by PoS. This information could be particularly interesting for researchers interested in studying the processing of different grammatical categories and/or in studying processing beyond the single word level (see Brysbaert et al., 2012; Hofmann et al., 2007).

Finally, the morpho-syntactic field also displays information concerning grammatical gender (masculine: e.g., *carro* [car]; feminine: e.g., *mesa* [table]; or both (fixed): e.g., *estudante* [student]), grammatical number (singular: e.g., *casa* [house]; plural: e.g., *casas* [houses]; or both (fixed): e.g., *lápis* [pencil]), and whether a given lexical entry stems from a foreign language (0 = false, 1 = true). Note that although such

words as "timing" or "briefing" are indexed as lexical entries in the P-PAL databases due to their widespread use in the EP language, they are not considered in the computation of the statistics provided in P-PAL, since they do not conform to the EP orthographic rules. Only loan words that have already been adapted to EP (e.g., *abajour* [lamp], *acordeão* [accordion], *anoraque* [anorak]) and that constitute lexical entries in the reference dictionaries (e.g., Casteleiro, 2001) are included in these computations (see Soares et al., 2014, for details).

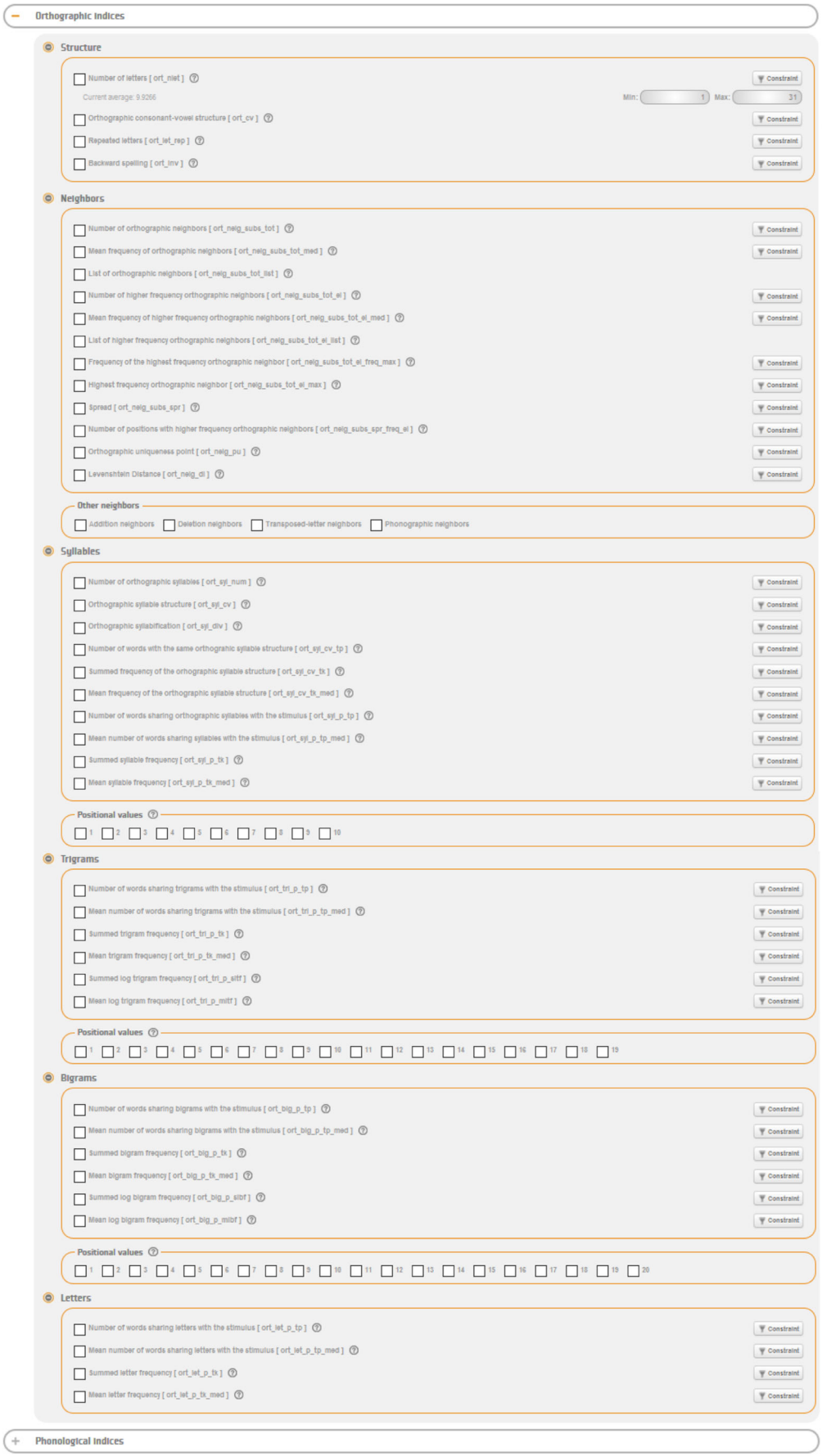
Orthographic statistics

This field integrates a broad range of orthographic attributes and lexical and sublexical statistics of progressively smaller grain sizes, as in other international databases (e.g., Balota et al., 2007; Boudelaa & Marslen-Wilson, 2010; Chetail & Mathey, 2010; Davis, 2005; Davis & Perea, 2005; Duchon et al., 2013; Duñabeitia et al., 2010; Hofmann et al., 2007; Ktori et al., 2008; Kyparissiadis et al., 2017; New et al., 2004; Perea et al., 2006). As is depicted in Fig. 7, this information is distributed into six subfields: orthographic structure information, orthographic neighborhood statistics, orthographic syllabic measures, and trigram, bigram, and letter frequency distributions.

Orthographic structure The word attributes displayed in this subfield are word length in number of letters [ort_nlet], word consonant [c]–vowel [v] orthographic structure [ort_cv], number of letters that occur more than once within the word [ort_let_rep], and the graphic representation of the backward spelling of the word [ort_inv]. Thus, for the EP word *casa* [house], for instance, P-PAL returns the following information: ort_nlet = 4, ort_cv = CVCV, ort_let_rep = a, and ort_inv = asac, both in the lemma and word form databases.

Orthographic neighborhood statistics P-PAL provides several measures regarding the number, distribution, and characteristics of the orthographic neighborhood for each of its lexical entries in the lemma and word form databases (see Fig. 8). Specifically, P-PAL provides Coltheart et al.'s (1977) classic *N* neighborhood metric, indexing all the words that can be formed by replacing a single letter at any position within the string, while maintaining the remaining letters in the same positions (ort_neig_subs_tot). It also provides the mean word frequency (ort_neig_subs_tot_med) and the list (ort_neig_subs_tot_list) of words that integrate that neighborhood, as well as the number (ort_neig_subs_tot_el), mean frequency (ort_neig_subs_tot_el_med), and list (ort_neig_subs_tot_el_list) of the higher-frequency orthographic neighbors. The highest-frequency orthographic neighbor (ort_neig_subs_tot_el_max) of a given lexical entry and its frequency value (ort_neig_subs_tot_el_freq_max) are also provided. Information regarding the number of positions at which orthographic substitution neighbors can be formed (ort_neig_subs_spr), a metric known as *Spread* (see Johnson & Pugh, 1994; Mathey & Zagar, 2000), and the number of positions from which higher-frequency orthographic substitution neighbors are derived (ort_neig_subs_spr_freq_el), are also available. The letter position, counting from the left, at which a word becomes distinguishable from its neighbors (orth_uniq_point)—that is, the word orthographic uniqueness point (OUP; e.g., Kwantes & Mewhort, 1999; Miller, Juhasz, & Rayner, 2006)—can also be obtained from the application. For example, because *casa* has *caso* [case] as an orthographic neighbor, P-PAL returns orth_uniq_point = 4, which indicates that only at Position 4 does *casa* become

Fig. 7 Depiction of the six orthographic subfields and the five phonological subfields displayed in the P-PAL Web-based interface.


Orthographic indices
Structure
Number of letters [ort_nlet]
Constraint
Current average: 9.9266
Min: 1 Max: 31
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_ei]
Mean frequency of higher frequency orthographic neighbors [ort_neig_subs_tot_ei_med]
List of higher frequency orthographic neighbors [ort_neig_subs_tot_ei_list]
Frequency of the highest frequency orthographic neighbor [ort_neig_subs_tot_ei_freq_max]
Highest frequency orthographic neighbor [ort_neig_subs_tot_ei_max]
Spread [ort_neig_subs_spr]
Number of positions with higher frequency orthographic neighbors [ort_neig_subs_spr_freq_ei]
Orthographic uniqueness point [ort_neig_pu]
Levenshtein Distance [ort_neig_dl]
Other neighbors
Addition neighbors
Deletion neighbors
Transposed-letter neighbors
Phonographic neighbors
Syllables
Number of orthographic syllables [ort_syl_num]
Orthographic syllable structure [ort_syl_cv]
Orthographic syllabification [ort_syl_div]
Number of words with the same orthographic syllable structure [ort_syl_cv_tp]
Summed frequency of the orthographic syllable structure [ort_syl_cv_tk]
Mean frequency of the orthographic syllable structure [ort_syl_cv_tk_med]
Number of words sharing orthographic syllables with the stimulus [ort_syl_p_tp]
Mean number of words sharing syllables with the stimulus [ort_syl_p_tp_med]
Summed syllable frequency [ort_syl_p_tk]
Mean syllable frequency [ort_syl_p_tk_med]
Positional values
1 2 3 4 5 6 7 8 9 10
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]
Summed log trigram frequency [ort_tri_p_sltf]
Mean log trigram frequency [ort_tri_p_mltf]
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]
Summed log bigram frequency [ort_big_p_slbf]
Mean log bigram frequency [ort_big_p_mlbf]
Positional values
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Letters
Number of words sharing letters with the stimulus [ort_let_p_tp]
Mean number of words sharing letters with the stimulus [ort_let_p_tp_med]
Summed letter frequency [ort_let_p_tk]
Mean letter frequency [ort_let_p_tk_med]
Phonological indices

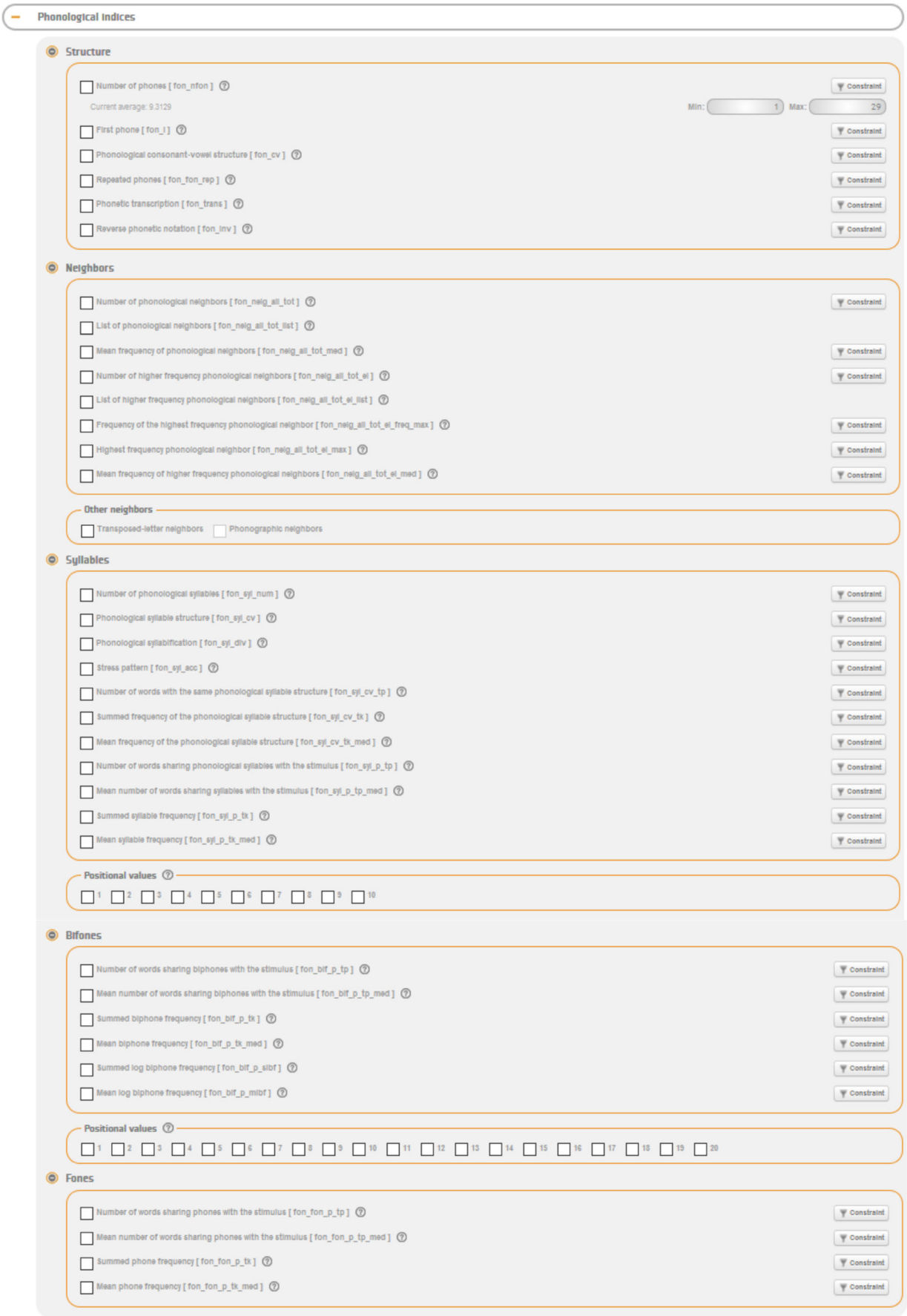
◀ **Fig. 8** Depiction of the measures available in each of the six orthographic subfields of the P-PAL Web-based interface. Note that this illustrates a "generate word list" query in the application, but the same statistics can be observed for the "analyze word list" query, except that the constraint options are not presented.

unique or can it be unequivocally identified in the lexicon. Moreover, the recent OLD20 neighborhood measure (Yarkoni et al., 2008) indexing the mean number of operations (i.e., substitutions, additions, and deletions) necessary to transform one word into another considering the 20 closest orthographic neighbors, is also presented. OLD 20 is assumed to be a rich and more flexible way of measuring orthographic similarity, as it deals with the negative relationship between word length and the number of orthographic neighbors more efficiently. Moreover, it accounts for higher percentages of variance, in both lexical decision and pronunciation performance, than the *N* metric (Yarkoni et al., 2008). Note that, unlike the *N* metric, which will be larger as the number of words in the neighborhood increases, in the OLD20 metric this value will be smaller, the larger the number of neighbors. For example, in the P-PAL word form database, the *N* value of *casa* is 22, indicating that *casa* has 22 neighbors created by replacing a single letter (i.e., *casa* has a dense neighborhood), whereas its OLD20 value is 1, which indicates that it takes only one operation to transform *casa* into each of its 20 closest neighbors (e.g., *caso*).

Furthermore, from the orthographic neighborhood statistics it is also possible to obtain the number, mean frequency, and list of other kinds of orthographic neighbors. Specifically, P-PAL provides these measures for the orthographic neighbors created by adding (orthographic addition neighbors; e.g., *casas* into *causas* [causes]) or by deleting (orthographic deletion neighbors; e.g., *casas* into *asas* [wings]) one single letter from the stimulus, as well as neighbors created by transposing two adjacent letters within the stimulus (orthographic transposition neighbors; e.g., *casas* into *cassas* [you revoke]) (see Davis & Andrews, 2001, or Davis, Perea, & Acha, 2009). Additionally, (substitution) neighbors that are simultaneously orthographic and phonological in nature, a type of neighborhood proposed by Peereman and Content (1997) as *phonographic neighbors*, are also presented in the application (see also Adelman & Brown, 2007). For instance, both *caso* ['kazu] and *cash* ['kɛ] are orthographic neighbors of *casa*, but only *caso* is a phonographic neighbor, since it takes more than one operation to transform *casa* ['kaz] into *cash* ['kɛ]. Finally, it is worth noting that since EP spelling makes use of different diacritics (e.g., ç á, à, â, ã, é, í, ó, õ, ú, ê, ô) that change both the visual form of the word and its pronunciation (e.g., the cedilla indicates that <ç> is pronounced [s] and not [k], as in <c>), affecting word processing (see Hermena, Liversedge, & Drieghe, 2016, for a recent eyetracking study showing diacritic effects on reading), the orthographic statistics provided in P-PAL take diacritics into account. Therefore, words such as *avô* [grandfather] and *avó* [grandmother] are considered orthographic neighbors.

Orthographic syllable measures The orthographic syllable measures provided in P-PAL include such syllabic attributes as the number of orthographic syllables within the word (ort_syl_num) (e.g., *casa* [house] has two orthographic syllables), the syllabified orthographic C–V structure (ort_syl_cv) of the word (e.g., *casa* presents a CV–CV orthographic syllable structure), and the orthographic syllabification according to the phonotactic and hyphenation rules of EP (ort_syl_div) (e.g., the orthographic syllabification of *casa* is ca–sa; see Fig. 9). Note, however, that although there is a match between the orthographic and phonological syllabifications of *casa*, there are cases in EP in which the orthographic and phonological syllabifications differ. For instance, in EP the double consonants <rr> and <ss> correspond to a single phoneme (/ʀ/ and /s/, respectively). Hence, words such as *carro* [car] and *pássaro* [bird] are phonetically syllabified into ['ka.ʀu] and ['pa.s . u] and orthographically syllabified into <car-ro> and <pás-sa-ro>. Moreover, there are also cases in which a word can be, for example, a disyllable in print and a monosyllable in speech, as in the EP word *leite* [milk]: <lei-te> versus [l jt]. Therefore, differences between the orthographic and phonological syllabifications in EP are to be expected.

Moreover, in line with the syllabic information provided in recent databases (e.g., Bédard et al., 2017; Chetail & Mathey, 2010; Davis, 2005; Davis & Perea, 2005; Duchon et al., 2013; Duñabeitia et al., 2010; Kyparissiadis et al., 2017), P-PAL also allows researchers to obtain several type and token positional syllable statistics. Specifically, it is possible to obtain the number (ort_syl_cv_tp), the summed word frequency (ort_syl_cv_tk), and the mean word frequency (ort_syl_cv_tk) of the words sharing the same syllable structure with the stimulus, as well as the number (ort_syl_p_tp), the mean (ort_syl_p_tp_med), the summed word frequency (ort_syl_p_tk), and the mean word frequency (ort_syl_p_tk_med) of the words with the same number of syllables sharing the same syllables in the same positions. For instance, in the P-PAL word form database contains 1,197 words (type frequency) with the same CV–CV syllabic structure as the word *casa* (e.g., *lata* [metal can], *bota* [boot]), totaling a summed word frequency (token frequency) and a mean token frequency of 47,257.6348 and 39.4801 pmw,


Phonological indices
Structure
Number of phones [fon_nfon]
Constraint
Current average: 9.3129
Min: 1 Max: 29
First phone [fon_l]
Phonological consonant-vowel structure [fon_cv]
Repeated phones [fon_fon_rep]
Phonetic transcription [fon_trans]
Reverse phonetic notation [fon_inv]
Neighbors
Number of phonological neighbors [fon_neig_all_tot]
List of phonological neighbors [fon_neig_all_tot_list]
Mean frequency of phonological neighbors [fon_neig_all_tot_med]
Number of higher frequency phonological neighbors [fon_neig_all_tot_ei]
List of higher frequency phonological neighbors [fon_neig_all_tot_ei_list]
Frequency of the highest frequency phonological neighbor [fon_neig_all_tot_ei_freq_max]
Highest frequency phonological neighbor [fon_neig_all_tot_ei_max]
Mean frequency of higher frequency phonological neighbors [fon_neig_all_tot_ei_med]
Other neighbors
Transposed-letter neighbors
Phonographic neighbors
Syllables
Number of phonological syllables [fon_syl_num]
Phonological syllable structure [fon_syl_cv]
Phonological syllabification [fon_syl_cliv]
Stress pattern [fon_syl_acc]
Number of words with the same phonological syllable structure [fon_syl_cv_tp]
Summed frequency of the phonological syllable structure [fon_syl_cv_tk]
Mean frequency of the phonological syllable structure [fon_syl_cv_tk_med]
Number of words sharing phonological syllables with the stimulus [fon_syl_p_tp]
Mean number of words sharing syllables with the stimulus [fon_syl_p_tp_med]
Summed syllable frequency [fon_syl_p_tk]
Mean syllable frequency [fon_syl_p_tk_med]
Positional values
1 2 3 4 5 6 7 8 9 10
Biphones
Number of words sharing biphones with the stimulus [fon_bif_p_tp]
Mean number of words sharing biphones with the stimulus [fon_bif_p_tp_med]
Summed biphone frequency [fon_bif_p_tk]
Mean biphone frequency [fon_bif_p_tk_med]
Summed log biphone frequency [fon_bif_p_slbf]
Mean log biphone frequency [fon_bif_p_mlbf]
Positional values
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Fones
Number of words sharing phones with the stimulus [fon_fon_p_tp]
Mean number of words sharing phones with the stimulus [fon_fon_p_tp_med]
Summed phone frequency [fon_fon_p_tk]
Mean phone frequency [fon_fon_p_tk_med]

respectively. Additionally, *casa* presents an ort_syl_p_tp = 280, showing that there are 280 words in the word form database with the same number of orthographic syllables (two) that share either the syllable <ca> at Position 1 or the syllable <sa> at Position 2, with ort_syl_p_tp_med = 140 indicating the mean number of words per syllable. Moreover, for the word *casa*, P-PAL returns ort_syl_p_tk = 5,843.6478, showing the summed pmw frequency of all the words (N = 280) that share syllables with the stimulus in the same positions, and ort_syl_p_tk_med = 20.8702, indicating the mean pmw frequency of these 280 words.

Finally, it is also possible to obtain type and token syllable statistics for specific syllables within the word, and not only for the entire word as in previous statistics. Following the same example, for the word *casa* it is possible to obtain the number (ort_syl_p_tp_1) and the summed frequency (ort_syl_p_tk_1) of the words with the same number of syllables that share the syllable <ca> at Position 1, as well as the number (ort_syl_ip_tp_1) and the summed frequency (ort_syl_ip_tk_1) of the words with the same number of syllables sharing syllable <ca> in any position (Position 1 or 2). The returned values for the syllable <ca> in casa at Position 1 are ort_syl_p_tp_1 = 185, and ort_syl_p_tk_1 = 2,992.1096, and at any position are ort_syl_ip_tp_1 = 308, and ort_syl_ip_tk_1 = 4,860.8041. Positional and nonpositional syllabic statistics for specific syllables within the word are provided from Position (Syllable) 1 to Position (Syllable) 10, because almost the entire lexicons (99%) in the lemma and word form databases have a syllable length below that number.

Trigram, bigram, and letter frequency distributions P-PAL also provides (type and token) statistics regarding the occurrence of trigrams (three-letter co-occurrences within the string; e.g., in the word *casa*, <cas> corresponds to Trigram 1, and <asa> to Trigram 2), bigrams (two-letter co-occurrences within the string; e.g., in the word *casa*, <ca> corresponds to Bigram 1, <as> to Bigram 2, and <sa> to Bigram 3) and letters (in the word *casa*, occurrences of <c>, <a>, <s>, and <a>), due to the relevance of these sublexical units in current visual word recognition research (e.g., Hand, O'Donnell, & Sereno, 2012; New & Grainger, 2011; see Fig. 9). Specifically, P-PAL provides statistics concerning the number (ort_tri_p_tp, ort_big_p_tp), the mean number (ort_tri_p_tp_med, ort_big_p_tp_med), the summed frequency (ort_tri_p_tk, ort_big_p_tk), the mean frequency (ort_tri_p_tk_med, ort_big_p_tk_med), the $\log_{10}$ transformation of the summed frequency (ort_tri_p_sltf, ort_big_p_sltf), and the $\log_{10}$ transformation of the mean frequency (ort_tri_p_mltf, ort_big_p_mltf) for both trigrams and bigrams, considering the whole string. All these measures are also length- and position-sensitive, as in other lexical databases (e.g., Balota et al., 2007; Boudelaa & Marslen-Wilson, 2010; Davis, 2005; Davis & Perea, 2005; Duchon et al., 2013; Hofmann et al., 2007; Ktori et al., 2008; Kyparissiadis et al., 2017; New et al. 2004). For an illustration, P-PAL returns the following trigram statistics for the word *casa*: ort_tri_p_tp = 7, indicating that 7 four-letter words in the word form database share the first, <cas>, or the second, <asa>, trigram with the stimulus; ort_tri_p_tp_med = 3.5, showing the mean number of words sharing the same trigrams with *casa* per trigram; ort_tri_p_tk = 1,522.836, indicating the summed pmw frequency of these words; ort_tri_p_tk_med = 217.548, showing the mean pmw frequency of these words; ort_tri_p_sltf = 5.6676, indicating the value corresponding to the $\log_{10}$ transformation of the summed pmw frequency of the words sharing the same trigrams with *casa*; and ort_tri_p_mltf = 2.8338, showing the $\log_{10}$ transformation of the mean pmw frequency of the words sharing the same trigrams. Similar statistics are provided for bigrams, although in this case a higher number of segments were computed. In addition, it is also possible to obtain positional and nonpositional statistics for specific trigrams and bigrams within the word, as for the syllabic statistics described above. For instance, P-PAL provides the type and token frequencies for the words with the same number of letters sharing a given trigram (or bigram) with the stimulus in a given position (e.g., Trigram 1 in Position 1), as well as the type and token frequencies for words with the same number of letters sharing a given trigram (e.g., Trigram 1) in any position of the string. These positional and nonpositional statistics are available from Position 1 to Position 19 for trigrams, and from Position 1 to Position 20 for bigrams, because they cover almost the entire P-PAL lexicon in the lemma and word form databases. As an illustration, *casa* shares Trigram 1 at Position 1 with four words in the word form lexicon (ort_tri_p_tp_1 = 4), with five words at any position (ort_tri_ip_tp_1 = 5), and presents a summed frequency of ~1,100 pmw, both for the words sharing Trigram 1 in Position 1 (ort_tri_p_tk_1 = 1,099.912) and for the words sharing Trigram 1 in any position (ort_tri_ip_tk_1 = 1,100.552). Finally, the letter statistics in P-PAL include the number (ort_let_p_tp) and mean number (ort_let_p_tp_med) of words with the same number of letters sharing the same letters in the same positions, as well as their summed (ort_let_p_tk) and mean (ort_let_p_tk_med) frequencies. Thus, the letter statistics obtained for *casa* show that there are ~1,200

◀ **Fig. 9** Depiction of the measures available in each of the five phonological subfields of the P-PAL Web-based interface. Note that this illustrates a "generate word list" query in the application, but the same statistics can be observed for the "analyze word list" query, except that the constraint options are not presented.

four-letter words sharing letter <c> in Position 1, letter <a> in Position 2, letter <s> in Position 3, and letter <a> in Position 4 (ort_let_p_tp = 1,180) in the word form database, and that the mean number of words per position (letter) is thus ~300 words (ort_let_p_tp_med = 295). The summed pmw frequency (ort_let_p_tk) of all these words is 6,2154.9195, and the mean pmw frequency (ort_let_p_tk_med) is 52.6737.

Phonological statistics

The phonological statistics provided in P-PAL mimic those presented in the orthographic field. They include a broad range of phonological attributes and lexical and sublexical statistics of different (decreasing) grain sizes (for the spoken word as a whole and for phonological syllables, biphones, and phones; see Fig. 7), in line with the phonological metrics available in other databases (e.g., Balota et al., 2007; Bédard et al., 2017; Chetail & Mathey, 2010; Davis, 2005; Davis & Perea, 2005; Duchon et al., 2013; Hofmann et al., 2007; Kyparissiadis et al., 2017; New et al., 2004; New & Spinelli, 2013). Note that the phonological information provided in P-PAL results from the phonetic transcription of all its lexical entries using the International Phonetic Alphabet (IPA) and from the computation of the different phonological statistics on the basis of the distributions observed in the lemma and word form corpus previously described (see the Corpus Sampling section), from which the orthographic statistics were also obtained. Nevertheless, it is worth noting that since EP is an intermediate-depth, stress-timed language in which the spelling-to-sound correspondences are not direct (see Campos, Mendes Oliveira, & Soares, 2018), differences between the orthographic and phonological statistics are to be expected, and researchers interested in studying the processes and mechanisms involved in the visual word recognition and reading of EP words should account for them in research, as we have mentioned. Below, the attributes and statistics provided in each of the five phonological subfields depicted in Fig. 7 are described.

Phonological structure The phonological information presented in this subfield includes, as in other databases (e.g., Balota et al., 2007; Chetail & Mathey, 2010; Davis, 2005; Davis & Perea, 2005; Duchon et al., 2013; Hofmann et al., 2007; Kyparissiadis et al., 2017; New et al., 2004; New & Spinelli, 2013), word properties such as the number of phones or phonemes (fon_nfon),[2] the first phone (fon_i), the CV phonological structure (fon_cv), the number of phones that occur more than once within the word (ort_fon_rep), the pronunciation of the word according to the standard accent in EP using the IPA phonetic symbols (fon_trans), and the reverse phonetic transcription of the word (fon_inv). Thus, for the spoken form of the EP word *casa* [house], for instance, P-PAL returns the following information: fon_trans = ['kaz], ort_inv = [zak'] (the diacritc marks the stress pattern of the word, indicating, in this case, that the first syllable is stressed), fon_nlet = 4, fon_i = k, fon_cv = CVCV, and ort_let_rep = empty (since no phone is repeated in the word; note that the two <a>s correspond to different EP vocalic sounds).

Phonological neighborhood statistics As in the ortographic field, P-PAL provides several measures regarding the distribution and the characteristics of the phonological neighborhood of each of its lexical entries (see Fig. 9). Specifically, in this subfield, P-PAL provides the number of phonological neighbors (fon_neig_all_tot)—that is, the number of words that differ from another word on the basis of a single phoneme that is either substituted, deleted, or added, following the classic proposal of Luce and Pisoni (1998)—as well as the mean word frequency (fon_neig_all_tot_med) and the list (fon_neig_all_tot_list) of the words that integrate the phonological neighborhood of a given word. For example, *casa* presents 19 phonological neighbors (e.g., ['ka], ['kaz], ['baz], [' az]) that present a mean word frequency of 42.503 pmw. Moreover, the number (fon_neig_all_tot_el) and list (fon_neig_all_tot_el_list) of phonological neighbors with a higher frequency are also provided, as well as their mean word frequency (fon_neig_all_tot_el_med). As with the orthographic neighborhood statistics, it is also possible to access the frequency (fon_neig_all_tot_el_freq_max) and the word that corresponds to the highest-frequency phonological neighbor of a given word (ort_neig_all_tot_el_max). For example, the highest-frequency phonological neighbor of ['kaz] is ['kazu], presenting a frequency of 676.4594 pmw. Finally, from this phonological subfield it is also possible to obtain the number and the list of phonological neighbors created by transposing two adjacent phones in the stimulus (phonological transposition neighbors, see Davis & Andrews, 2001, or Davis et al., 2009), as well as their mean frequency. Although in its spoken form *casa* has no transposition neighbors, the spoken form of the word *abril* [april], for instance, presents *baril* [cool] as its phonological transposition neighbor, in both the lemma and word form P-PAL databases. The access to phonographic neighbors of a given word (i.e., neighbors that are simultaneously orthographic and

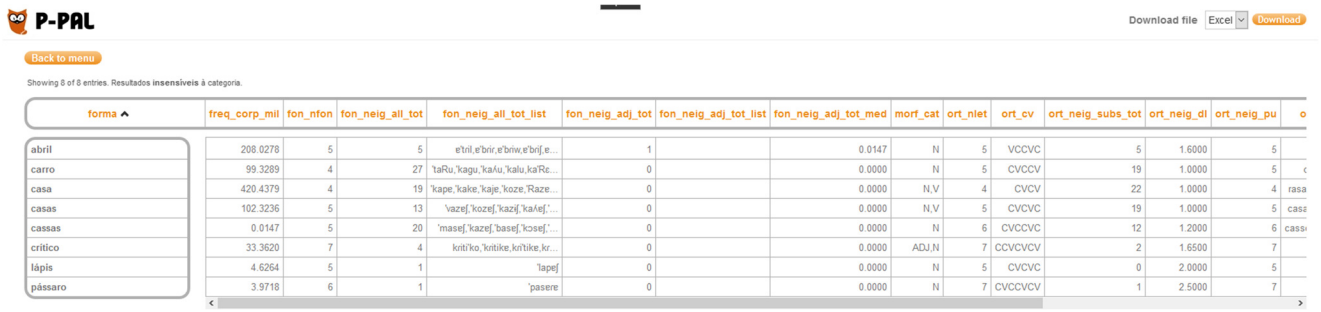
[2] Note that since we computed all the phonological measures provided in P-PAL on the basis of the IPA phonetic transcriptions of all its lexical entries, we opted to use the term "phones" instead of "phonemes" in the application, but both terms can be used, since the phonetic transcriptions used captured the contrastive sounds that are meaningful in the EP language.

phonological, see Adelman & Brown, 2007; Peereman & Content, 1997) is also allowed from this subfield of the application.

Phonological syllable measures The phonological syllable attributes and statistics provided in P-PAL include the number of phonological syllables in the word (fon_syl_num) (e.g., in its spoken form casa has two phonological syllables), the phonological syllabification of the word according to the standard accent in EP (fon_syl_div) (e.g., the phonological syllabification of casa is ['ka.z]), the phonological CV structure of the word (fon_syl_cv) (e.g., the phonological syllable structure of *casa* is CV.CV), and the stress pattern of the word, using 1 for the position in which the syllable is stressed and 0 for the unstressed syllable positions (fon_syl_acc) (e.g., the stress pattern of *casa* is "1.0," indicating that the first syllable is the stressed one). Besides these attributes, P-PAL also offers, in line with the syllabic information provided in recent databases (e.g., Bédard et al., 2017; Chetail & Mathey, 2010; Davis, 2005; Davis & Perea, 2005; Duchon et al., 2013; Duñabeitia et al., 2010; Kyparissiadis et al., 2017; New & Spinelli, 2013), several statistics regarding the number of words sharing the same phonological syllable structure as the stimulus (fon_syl_cv_tp) (see Fig. 9), as well as their summed (fon_syl_cv_tk) and mean (fon_syl_cv_tk) word frequencies, thus mimicking the statistics provided in the orthographic syllable subfield. Specifically, from the phonological syllable statistics, it is possible to observe, for example, that the spoken form of the word casa has 2,344 words sharing the same CV.CV phonological syllable structure in the word form database (note that in the written form it only shared the orthographic syllable structure with 1,197 words), totaling a summed frequency and mean frequency of 64,687.9225 and 27.5972 pmw, respectively. Moreover, as for the orthographic syllable measures, P-PAL also provides the number (fon_syl_p_tp) and mean number (fon_syl_p_tp_med) of words with the same number of phonological syllables containing the same syllables in the same positions, as well as their summed (fon_syl_p_tk) and mean (fon_syl_p_tk_med) word frequencies. For instance, the returned statistics for the spoken form of *casa* are fon_syl_p_tp = 167, indicating that there are 167 words in the word form database that share either the first [ka] or the second [z] phonological syllable with the stimulus; fon_syl_p_tp_med = 83.5, showing the mean number of words per phonological syllable; fon_syl_p_tk = 3,279.4808, indicating the summed pmw frequency of these 167 words; and fon_syl_p_tk_med = 19.6376, showing their mean pmw frequency.

Finally, it is also possible to obtain syllabic statistics for specific phonological syllables within the stimulus—that is, the number (fon_syl_p_tp_1) and the summed frequencies (fon_syl_p_tk_1) of words with the same number of phonological syllables that share a given phonological syllable (e.g., [ka]) in a given position (e.g., first), as well as the number (fon_syl_ip_tp_1) and the summed frequencies (fon_syl_ip_tk_1) of words with the same number of phonological syllables sharing a given phonological syllable (e.g., [ka]) in any position (in this case, in the first or second positions). For example, the returned statistics for the first syllable of *casa* are fon_syl_p_tp_1 = 103, indicating that 103 words in the word form database share the syllable [ka] in first position; fon_syl_p_tk_1 = 1,972.6209, showing the summed frequency of these 103 words; fon_syl_ip_tp_1 = 105, indicating that two more words in the word form database shared the syllable [ka] when also considering the second position; and fon_syl_ip_tk_1 = 1,972.6356, showing the summed frequency of these 105 words. The positional and nonpositional phonological syllable statistics are provided from Syllable 1 to Syllable 10, as in the case of the orthographic syllable measures.

Biphone and phone frequency distributions In this phonological subfield, P-PAL provides the type and the token frequency distributions of biphones (i.e., co-occurrences of two phones within the string; e.g., in the word *casa* [ka] corresponds to Biphone 1, [az] to Biphone 2, and [z] to Biphone 3), and phones (i.e., occurrences of [k], [a], [z], and [ ], as in the spoken form of the word *casa*), for each of the lexical entries in the lemma and word form databases. The biphone and phone statistics are length- and position-sensitive, like the equivalent grain size statistics presented in the orthographic field. Specifically, P-PAL provides biphone statistics targeting the number (fon_bif_p_tp) and mean number (fon_bif_p_tp_med) of words sharing the same biphones with the stimulus in the same positions considering the entire string, their summed (fon_bif_p_tk) and mean (fon_bif_p_tk_med) frequencies, the $\log_{10}$ transformation of the summed frequency (fon_bif_p_sltf), and the $\log_{10}$ transformation of the mean frequency (fon_fif_p_mltf), as well as statistics for specific dual units within the string, such as the number of words with the same number of phones sharing a given biphone in a given position (e.g., Biphone 1 at Position 1, fon_bif_p_tp_1) or at any position (e.g., Biphone 1 at Position 1 and Position 2, fon_bif_ip_tp_1) and its corresponding summed frequencies (fon_bif_p_tk_1, fon_bif_ip_tk_1, respectively). For example, the returned biphone statistics for the spoken form of the word *casa* are fon_bif_p_tp = 102, indicating that there are 102 words with four phonemes sharing the first [ka],



forma	freq_corp_mil	fon_nfon	fon_neig_all_tot	fon_neig_all_tot_list	fon_neig_adj_tot	fon_neig_adj_tot_list	fon_neig_adj_tot_med	morf_cat	ort_nlet	ort_cv	ort_neig_subs_tot	ort_neig_dl	ort_neig_pu	o
abril	208.0278	5	5	e'tril,e'brir,e'briw,e'brif,e...	1		0.0147	N	5	VCCVC	5	1.6000	5	
carro	99.3289	4	27	'taRu,'kagu,'ka4u,'kalu,ka'Rɛ...	0		0.0000	N	5	CVCCV	19	1.0000	5	c
casa	420.4379	4	19	'kape,'kake,'kaje,'koze,'Raze...	0		0.0000	N,V	4	CVCV	22	1.0000	4	rasa
casas	102.3236	5	13	'vazeʃ,'kozeʃ,'kaziʃ,'kaʌeʃ,'...	0		0.0000	N,V	5	CVCVC	19	1.0000	5	casa
cassas	0.0147	5	20	'maseʃ,'kazeʃ,'baseʃ,'kɔseʃ,'...	0		0.0000	N	6	CVCCVC	12	1.2000	6	cass
crítico	33.3620	7	4	kriti'ko,'kritike,kri'tike,kr...	0		0.0000	ADJ,N	7	CCVCVCV	2	1.6500	7	
lápis	4.6264	5	1	'lapeʃ	0		0.0000	N	5	CVCVC	0	2.0000	5	
pássaro	3.9718	6	1	'pasere	0		0.0000	N	7	CVCCVCV	1	2.5000	7	

Fig. 10 Depiction of the online output file in the P-PAL Web-based interface.

the second [az], and/or the third [z] biphone with the stimulus, in the same positions; fon_bif_p_tp_med = 34, showing the mean number of words per biphone in the word; fon_bif_p_tk = 3,678.4613, indicating the pmw summed frequency of the 34 words sharing the same biphones at the same positions; fon_bif_p_tk_med = 36.0633, showing the mean pmw frequency of these words; fon_bif_p_sltf = 9.1575, the $\log_{10}$ transformation of the pmw summed frequency value; and fon_bif_p_mltf = 3.0525, indicating the $\log_{10}$ transformation of the mean pmw frequency of the words sharing the same biphones in the same positions. Moreover, the returned statistics for specific biphones show that *casa* shares Biphone 1 ([ka]) at Position 1 with other 51 words in the word form database (fon_bif_p_tp_1 = 51) and presents a summed frequency of ~1,500 occurrences pmw (fon_bif_p_tk_1 = 1,533.9112). Because in four-phone words the biphone [ka] only occurs in Position 1, the same values are returned when the user asks for nonpositional biphone statistics (fon_bif_ip_tp_1 = 51, fon_bif_ip_tk_1 = 1,533.9112). Statistics for specific biphones are available from Positions 1 to 20, as for the bigram statistics. Finally, the phone statistics provided in P-PAL mimic those presented for letters in the orthographic field, and include measures such as the number (fon_fon_p_tp) and mean number (fon_fon_p_tp_med) of words with the same number of phones as the stimulus that share the same phones in the same positions, as well as their summed (fon_fon_p_tk) and mean (fon_fon_p_tk_med) frequencies. Following the same example, P-PAL shows that *casa* presents 1,636 word forms with four phonemes that share phone [k] at Position 1, phone [a] at Position 2, phone [z] at Position 3, or phone [ ] at Position 4. The mean number of words per position is 409, and the summed and mean frequencies of these words are 5,8340.8086 and 35.6606 pmw, respectively.

Finally, it is worth noting that after the user has selected the word attributes and lexical and/or sublexical statistics, he or she can run the word query by clicking the "execute query" button displayed at the lower right corner of the analysis menu. The output is immediately displayed in the format presented in Fig. 10.

The output resembles a spreadsheet in which each word is presented vertically on a separate line and each attribute/statistic selected is shown horizontally in different columns. The user can analyze the output online and/or save it as an Excel file (.xls or .cvc) by clicking the "download" option displayed in the upper right corner of the output menu. This option is extremely useful, since it allows researchers to continue work offline using the Excel file. For instance, users can delete, filter, or combine the data provided in new ways according to the purposes of their research. Note, however, that P-PAL only returns word attributes and statistics for no more than 15,000 lexical entries (lemmas or word forms) with each word query. Researchers interested in longer word lists are encouraged to apply more and/or finer constraints to the word search. The output provided by P-PAL for the "analyze word list" option follows the same order as the input file uploaded, whereas for a "generate word list" query, words are displayed alphabetically.

Conclusion

In this work we have presented the procedures involved in the development of a new EP lexical database, P-PAL, which provides researchers with a broad range of word attributes and statistics not yet available for EP, including several measures of word frequency, morpho-syntactic information, as well as numerous lexical and sublexical orthographic and phonological statistics of different grain sizes (word as a whole, syllables, bigrams/biphones, and letters/phones) for ~53,000 lemmatized and ~208,000 nonlemmatized (word forms) EP words. These statistics were drawn from a large-size (over 227 million words) and diversified contemporary EP corpus (containing written and spoken records from different language resources and genres), in order to best represent the EP language and minimize error in the computation of these metrics. Moreover, we also present the Web-based interface developed to support this new EP lexical database and allow

researchers from different fields of study (e.g., psycholinguistics, linguistics, neurosciences, or cognitive psychology in general) to obtain EP word attributes and statistics in a quick and efficient way. The P-PAL Web-based interface combines two types of word queries in both the lemma and word form databases: (i) It can analyze words previously selected by the researcher for specific attributes and lexical and/or sublexical characteristics, and (ii) it can generate word lists that meet specific word requirements defined by the user in the menu of analysis. These word query options give strong versatility to the research tool and increase its usefulness in supporting well-controlled and well-designed research using EP verbal stimuli. In sum, for the potential it brings to research and the entirely new set of orthographic and phonological lexical and sublexical statistics it provides, P-PAL will be a key resource for the development and internationalization of the research with EP verbal stimuli. The P-PAL Web-based interface is freely available for research purposes at http://p-pal.di.uminho.pt/tools.

Author note This study was conducted at the Psychology Research Centre (PSI/01662), University of Minho. It was supported by the Portuguese Foundation for Science and Technology and the Portuguese Ministry of Science, Technology, and Higher Education through national funds, and co-financed by FEDER through European funds (Grant POCI-01-0145-FEDER-007653).

References

Adelman, J. S., & Brown, G. D. A. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review*, *14*, 455–459. https://doi.org/10.3758/BF03194088

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*, 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, *11*, 295–328.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, *37*, 94–117. https://doi.org/10.1006/jmla.1997.2509

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*, 290–313. doi: https://doi.org/10.1016/j.jml.2006.03.008

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). The CELEX Lexical Database (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Balota, D., Yap, M., & Cortese, M. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 285–375). Amsterdam, The Netherlands: Academic Press.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . .Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459. https://doi.org/10.3758/BF03193014

Bédard, P., Audet, A.-M., Drouin, P., Roy, J.-P., Rivard, J., & Tremblay, P. (2017). SyllabO+: A new tool to study sublexical phenomena in spoken Québec French. *Behavior Research Methods*, *49*, 1852–1863. https://doi.org/10.3758/s13428-016-0829-7

Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for Modern Standard Arabic. *Behavior Research Methods*, *42*, 481–487. https://doi.org/10.3758/BRM.42.2.481

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science*, *7*, 96–99.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*, 412–424. https://doi.org/10.1027/1618-3169/a000123

Brysbaert, M., & Diependaele, K. (2013). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods*, *45*, 422–430. https://doi.org/10.3758/s13428-012-0270-5

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. doi: https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*, 991–997. https://doi.org/10.3758/s13428-012-0190-4

Campos, A. D., Mendes Oliveira, H., & Soares, A. P. (2018). The role of syllables in intermediate-depth stress-timed languages: Masked priming evidence in European Portuguese. *Reading and Writing*, *31*, 1209–1229. https://doi.org/10.1007/s11145-018-9835-8

Casteleiro, J. M. (2001). *Dicionário da língua portuguesa contemporânea da Academia das Ciências de Lisboa* [Dictionary of the Portuguese contemporary language of the Lisbon Academy of Sciences]. Lisbon, Portugal: Academia das Ciências de Lisboa/ Editorial Verbo.

Chetail, F., & Mathey, S. (2010). InfoSyll: A syllabary providing statistical information on phonological and orthographic syllables. *Journal of Psycholinguistic Research*, *39*, 485–504.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505. https://doi.org/10.1080/14640748108400805

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, *37*, 65–70. https://doi.org/10.3758/BF03206399

Davis, C. J., & Andrews, S. (2001). Inhibitory effects of transposed-letter similarity for words and non-words of different lengths. *Australian Journal of Psychology*, *53*, 50.

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. https://doi.org/10.3758/BF03192738

Davis, C. J., Perea, M., & Acha, J. (2009). Re(de)fining the orthographic neighbourhood: The role of addition and deletion neighbours in lexical decision and reading. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1550–1570. https://doi.org/10.1037/a0014253

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, *45*, 1246–1258. https://doi.org/10.3758/s13428-013-0326-1

Duñabeitia, J. A., Cholin, J., Corral, J., Perea, M., & Carreiras, M. (2010). SYLLABARIUM: An online application for deriving complete statistics for Basque and Spanish orthographic syllables. *Behavior Research Methods*, *42*, 118–125. https://doi.org/10.3758/BRM.42.1.118

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, *28*, 1109–1115. https://doi.org/10.3758/BF03211812

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 104–115. https://doi.org/10.1037/0096-1523.13.1.104

Gomes, I., & Castro, S. L. (2003). Porlex, a lexical database in European Portuguese. *Psychologica*, *32*, 91–108.

Goswami, U., Ziegler, J. C., Dalton, L., & Schneider, W. (2001). Pseudohomophone effects and phonological recoding procedures in reading development in English and German. *Journal of Memory and Language*, *45*, 648–664. https://doi.org/10.1006/jmla.2001.2790

Grainger J., & Ziegler, J. C. (2011). A dual-route approach to orthographic processing. *Frontiers in Psychology*, *2*, 54. https://doi.org/10.3389/fpsyg.2011.00054

Grzybek, P. (2006). History and methodology of word length studies: The state of the art. In P. Grzybek (Ed.), *Contributions to the science of text and language: Word length studies and related issues* (pp. 15–90). Dordrecht, The Netherlands: Springer.

Hand, C. J., O'Donnell, P. J., & Sereno, S. C. (2012). Word-initial letters influence fixation durations during fluent reading. *Frontiers in Psychology*, *3*, 85:1–19. https://doi.org/10.3389/fpsyg.2012.00085

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB—Eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, *62*, 10–20. https://doi.org/10.1026/0033-3042/a000029

Hermena, E. W., Liversedge, S. P., & Drieghe, D. (2016). Parafoveal processing of Arabic diacritical marks. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 2021–2038.

Hofmann, M. J., Stenneken, P., Conrad, M., & Jacobs, A. (2007). Sublexical frequency measures for orthographic and phonological units in German. *Behavior Research Methods*, *39*, 620–629.

Johnson, N. E., & Pugh, K. R. (1994). A cohort model of visual word recognition. *Cognitive Psychology*, *26*, 240–346.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch words frequency based on film subtitles. *Behavior Research Methods*, *42*, 643–650. https://doi.org/10.3758/BRM.42.3.643

Ktori, M., van Heuven, W. J. B., & Pitchford, N. J. (2008). GreekLex: A lexical database of Modern Greek. Behavior Research Methods, 40, 773–783. https://doi.org/10.3758/BRM.40.3.773

Kwantes, P. J., & Mewhort, D. J. K. (1999). Evidence for sequential processing in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 376–381. https://doi.org/10.1037/0096-1523.25.2.376

Kyparissiadis, A., van Heuven, W. J. B., Pitchford, N. J., & Ledgeway, T. (2017). GreekLex 2: A comprehensive lexical database with part-of-speech, syllabic, phonological, and stress information. *PLoS ONE*, *12*, e0172493. https://doi.org/10.1371/journal.pone.0172493

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*, 1–36.

Mathey, S., & Zagar, D. (2000). The neighborhood distribution effect in visual word recognition: Words with single and twin neighbors. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 184–205. https://doi.org/10.1037/0096-1523.26.1.184

Miller, B., Juhasz, B. J., & Rayner, K. (2006). The orthographic uniqueness point and eye movements during reading. *British Journal of Psychology*, *97*, 191–216.

Nascimento, M. F. B., Marques M. L. G., & Cruz, M. L. S. (1987). *Português fundamental: Métodos e documentos. Vol. II, Tomo I: Inquérito de frequência* [Basic Portuguese: Methods and documents. Vol. II, Tomo I: Frequency survey]. Lisbon, Portugal: INIC, Centro de Linguística da Universidade de Lisboa.

Nascimento, M. F. B., Pereira, L. A. S., & Saramago, J. (2000). Portuguese Corpora at CLUL. In *Proceedings of the Second International Conference on Language Resources and Evaluation* (Vol. II, pp. 1603–1607). Athens, Greece: European Language Resources Association.

Nascimento, M. F. B., Rivenc, M. L. P., & Cruz, M. L. S. (1987). *Português fundamental: Métodos e documentos. Vol. II, Tomo II: Inquérito de disponibilidade* [Basic Portuguese: Methods and documents. Vol. II, Tomo II: Availability survey]. Lisbon, Portugal: INIC, Centro de Linguística da Universidade de Lisboa.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*, 661–677. https://doi.org/10.1017/S014271640707035X

New, B., & Grainger, J. (2011). On letter frequency effects. *Acta Psychologica*, *138*, 322–328. https://doi.org/10.1016/j.actpsy.2011.07.001

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*, 516–524. https://doi.org/10.3758/BF03195598

New, B., & Spinelli, E. (2013). Diphones-fr: A French database of diphones positional frequency. *Behavior Research Methods*, *45*, 758–764.

Parmentier, F. B. R., Comesaña, M., & Soares, A. P. (2017). Disentangling the effects of word frequency and contextual diversity on serial recall performance. *Quarterly Journal of Experimental Psychology*, *70*, 1–17. https://doi.org/10.1080/17470218.2015.1105268

Peereman, R., & Content, A. (1997). Orthographic and phonological neighborhoods in naming: Not all neighbors are equally influential in orthographic space. *Journal of Memory and Language*, *37*, 382–410.

Perea, M., Soares, A. P., & Comesaña, M. (2013). Contextual diversity is a main determinant of word-identification times in young readers. *Journal of Experimental Child Psychology*, *116*, 37–44.

Perea, M., Urkia, M., Davis, C. J., Agirre, A., Laseka, E., & Carreiras, M. (2006). E-Hitz: A word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). *Behavior Research Methods*, *38*, 610–615. https://doi.org/10.3758/BF03193893

Sebastián-Gallés, N., Martí, M. A., Cuetos, F., & Carreiras, M. (2000). *LEXESP: Una base de datos informatizado del español* (Spanish Computerized Lexicon). Barcelona, Spain: Ediciones de la Universitat de Barcelona.

Simões, A. M., & Almeida, J. J. (2001). Jspell: Um módulo de análise morfológica para uso em Processamento de Linguagem Natural. In A. Gonçalves & C. N. Correia (Eds.), Actas do Encontro Nacional da Associação Portuguesa de Linguística (pp. 485-495). Lisboa, Portugal: Associação Portuguesa de Linguística.

Sinclair, J. (2005). Corpus and text: Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxford, UK: Oxbow.

Soares, A. P., Iriarte, A., Almeida, J. J., Simões, A., Costa, A., França, P., … Comesaña, M. (2014). Procura-PALavras (P-PAL): Uma nova medida de frequência lexical do Português Europeu contemporâneo [Procura-PALavras (P-PAL): A new measure of word frequency for contemporary European Portuguese]. *Psicologia: Reflexão e Crítica*, *27*, 1–14.

Soares, A. P., Machado, J., Costa, A., Iriarte, A., Simões, A., de Almeida, J. J., … Perea, M. (2015). On the advantages of word frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. *Quarterly Journal of Experimental Psychology*, *68*, 680–696. https://doi.org/10.1080/17470218.2014.964271

Thorndike, E. L. (1921). *The teacher's word book*. New York, NY: Teachers College, Columbia University.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*, 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Yap, M., & Balota, D. (2015). Visual word recognition. In A. Pollastsek & R. Treiman (Eds.), *The Oxford handbook of reading* (pp. 26–43). New York, NY: Oxford University Press.

Yarkoni, T., Balota, D., & Yap, M. J. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971–979. https://doi.org/10.3758/PBR.15.5.971