


## Psycholinguistic variables in visual word recognition and pronunciation of European Portuguese words: a mega-study approach

Ana Paula Soares<sup>a</sup>, Alexandrina Lages<sup>a</sup>, Ana Silva<sup>a</sup>, Montserrat Comesaña<sup>a</sup>, Inês Sousa<sup>b</sup>, Ana P. Pinheiro <sup>c</sup> and Manuel Perea<sup>d,e</sup>

<sup>a</sup>CIPsi, Escola de Psicologia, Universidade do Minho, Braga, Portugal; <sup>b</sup>Departamento de Matemática e Aplicações, Universidade do Minho, Braga, Portugal; <sup>c</sup>Faculdade de Psicologia, Universidade de Lisboa, Lisbon, Portugal; <sup>d</sup>Departamento de Metodología, Universitat de València, Valencia, Spain; <sup>e</sup>Facultad de Lenguas y Educación, Universidad Nebrija, Madrid, Spain

### ABSTRACT

An increasing number of psycholinguistic studies have adopted a megastudy approach to explore the role that different variables play in the speed and/or accuracy with which words are recognised and/or pronounced in different languages. However, despite evidence for deep and shallow orthographies, little is known about the role that several orthographic, phonological and semantic variables play in visual word recognition and word production of words from intermediate-depth languages, as European Portuguese (EP). The current study aimed to overcome this gap, by collecting lexical decision and naming data for a large pool of words selected to closely represent the diversity of the EP language. Results from multiple regression analyses conducted on the latency data from both tasks place EP in-between the results previously observed in other deep- and shallow-orthographies. These findings indicate that EP represents a pivotal language to study the universality of the processes/mechanisms involved in skilled reading across languages.

### ARTICLE HISTORY

Received 22 May 2018  
Accepted 24 January 2019

### KEYWORDS

European Portuguese; megastudy; psycholinguistic variables; lexical decision; naming

### Introduction

Recent studies have provided new insights into the role played by different psycholinguistic variables in the speed and accuracy with which words are recognised and/or pronounced in different languages by adopting a megastudy approach. This approach presents several advantages over the classic factorial designs (e.g. see Balota, Yap, Hutchison, & Cortese, 2012; or Keuleers & Balota, 2015; for recent reviews). Indeed, since it involves the collection of reaction times and accuracy responses for a large amount of words using standard psycholinguistic tasks such as lexical decision and/or word naming (e.g. Balota et al., 2007; Ferrand et al., 2010; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2012; Yap, Liow, Jalil, & Faizal, 2010), experimenters do not need to match experimental stimuli on each of the potential variables that may affect the results (e.g. word frequency, word length, neighbourhood size). Instead, the influence of each word feature can be entered into the analyses (typically regression analysis) as predictors. Moreover, the variables under study do not need to be categorically classified (e.g. high vs. low frequency words and/or as short vs. long length words), as they can assume, in the statistical analyses, the “real” values that has been assigned to them

based on the information provided in lexical databases (e.g. per million word frequency, number of letters or syllables). These options increase the statistical power of the analyses to be conducted and the reliability of the results obtained as the effects observed are less dependent on the characteristics of the pool of stimuli (usually small) used, which often present extreme values in the dimensions under manipulation. Indeed, to maximise the odds of obtaining significant effects in factorial designs, researchers usually select the items that present the highest and the lowest values on a given dimension (e.g. word frequency, word length, neighbourhood size), which contributes to an over-representation of unusual items in the experimental list. Working with a large amount of items from a wide range of characteristics avoids this bias. Moreover, the megastudy approach also allows researchers to explore nonlinear effects on word processing (see, for instance, Baayen, Feldman, & Schreuder, 2006; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; or New, Ferrand, Pallier, & Brysbaert, 2006) and to reduce the influence of the experimenter in stimuli selection (see Forster, 2000).

Due to these advantages, it is not surprising that word processing times and accuracy rates have been collected

from an increasing number of large-scale studies conducted in different languages, such as English (e.g. Balota et al., 2004, 2007; Goh, Yap, Lau, Ng, & Tan, 2016; Keuleers et al., 2012; Yap & Balota, 2009), Italian (e.g. Barca, Burani, & Arduino, 2002; Burani, Arduino, & Barca, 2007), French (e.g. Ferrand et al., 2010, 2011, 2017), Dutch (e.g. Brysbaert, Stevens, Mandera, & Keuleers, 2016; Ernestus & Cutler, 2015; Keuleers et al., 2010), Spanish (e.g. Cuetos, Glez-Nosti, Barbon, & Brysbaert, 2011; Davies, Barbón, & Cuetos, 2013; González-Nosti, Barbón, Rodríguez-Ferreiro, & Cuetos, 2014; Wilson, Cuetos, Davies, & Burani, 2013), Malay (e.g. Yap et al., 2010), or Chinese (e.g. Sze, Rickard Liow, & Yap, 2014; Sze, Yap, & Rickard Liow, 2015; Tsang et al., 2017; Tse et al., 2017). Nonetheless, it is important to note that despite the renewed interest that conducting large-scale studies has been gathering in the scientific community in recent years, studies aiming to test which variables impacted strongly word processing from a large pool of items are not entirely new. In 1989, Seidenberg and Waters collected naming latencies for 2,897 monosyllabic English words. This study was followed by other works in English as well as in other alphabetic and non-alphabetic languages (e.g. Balota et al., 2004; Barca et al., 2002; Chateau & Jared, 2003; Cortese & Khanna, 2007; Cuetos & Barbón, 2006; Spieler & Balota, 2000; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995). Nevertheless, this approach was taken much further when Balota et al. (2007) published a chronometric database providing lexical decision and naming latencies for over 40,000 words and 40,000 non-words from American adult skilled readers, in what become known as the English Lexicon Project, then extended to other languages such as French (Ferrand et al., 2010), Dutch (Keuleers et al., 2010; see also Brysbaert et al., 2016 for an extension), Malay (Yap et al., 2010), English-British (Keuleers et al., 2012), and Chinese (Sze et al., 2014; see also Tsang et al., 2017; Tse et al., 2017).

Based on these chronometric datasets, several papers have been published revisiting factors that are known to affect word processing in different languages (e.g. word frequency, word length, syllables and morphemes, imageability and concreteness, Age-of-Acquisition [AoA], words' affective content – e.g. Bonin, Méot, & Bugaiska, 2018; Brysbaert et al., 2011; Cortese & Schock, 2013; Ferrand et al., 2010; Gimenes, Brysbaert, & New, 2016; Keuleers et al., 2010; Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011; Kuperman, 2015; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; New et al., 2006; Yap, Tan, Pexman, & Hargreaves, 2011). Other studies tested the role that new variables assume in word recognition as, for instance, the

Orthographic Levensthein Distance (OLD<sub>20</sub>) measure proposed by Yarkoni, Balota, and Yap (2008), or the word prevalence measure proposed recently by Brysbaert et al. (2016). Additionally, these datasets have also been used to compare the predictive power of different measures of the same variable (e.g. word frequency drawn from subtitles or from written-texts) on word processing (e.g. see Soares, Machado, et al., 2015 for a recent review), and/or to analyze the role that the same variables (e.g. word frequency, word length) played in visual word recognition and/or pronunciation of words from different languages (e.g. Ferrand et al., 2010; Gimenes et al., 2016; Keuleers et al., 2010, 2012; Yap et al., 2010).

This latter line of studies is particularly relevant as recent studies suggest that the processes and mechanisms involved in skilled reading are shaped by language characteristics. For instance, Sze et al. (2015), in a recent study aimed to analyze the role that different variables play in the speed with Chinese characters were recognised taking advantage of the lexical decision latencies provided by the Chinese Lexicon Project (Sze et al., 2014), showed that in a logographic language such as Chinese, the semantic and orthographic variables accounted for the highest percentage of variance, whereas the phonological variables contributed only for a modest percentage of variance (see Sze et al., 2015 for details). These results are remarkable not only because they shed light on the variables that affect word processing in a non-alphabetic language, but, importantly, because they challenged the accepted view that phonology is inevitable in reading, even in languages with a lack of correspondence between orthography and phonology such as Chinese (see Frost, 1998; Perfetti, Zhang, & Berent, 1992; Ziegler, Tan, Perry, & Montant, 2000). Hence, using chronometric datasets to explore the impact of the same variables across languages can be highly advantageous. It allows testing the universality of the processes/mechanisms involved in skilled reading across languages.

However, despite these datasets are available from a growing number of languages, they are still lacking for European Portuguese (EP). This is an important limitation since EP presents several orthographic and phonological characteristics that are distinct from deep and shallow orthographies. For example, EP presents an orthographic system that is more opaque than Spanish or Italian, though less opaque than English or French. It is therefore considered an intermediate-depth language which has been shown to have an important impact on reading acquisition (e.g. Fernandes, Ventura, Querido, & Morais, 2008; Seymour, Aro, & Erskine, 2003). Moreover, differently from Spanish, French or Italian, but similarly to

English, in EP, syllables tend not to follow each other at regular intervals, which might contribute to attenuate the role that syllables assume as a perceptual unit in word recognition. Indeed in EP, syllabic boundaries are less clearly defined than in the above cited languages, as they are blurred by phenomena such as vowel reduction (i.e. in spoken language, many vowels are not pronounced, causing mismatches between syllable divisions in speech and in print – for instance, the final vowel in the word *leite*[milk] is not pronounced, making *leite* to be a monosyllable in speech [ˈlɛjtɨ] and a dissyllable in print <lei-te>) and ambisyllabicity (i.e. a given consonant may work both as the coda of one syllable or the onset of the following syllable giving rise to different syllable divisions as in the word *acne*[acne] that can be divided both as /a.cne/ or /ac.ne/) (see Campos, Oliveira, & Soares, 2018; for more details). Furthermore, EP is also considered one of the Indo-European languages with higher syllabic and morphological complexity. In EP words can be created not only through the addition of prefixes and/or suffixes as in English, but also by compounding two or more morphemes into one single word. Consequently, words are longer in EP than in other Romance languages. According to the Procura-PALavras lexical database (P-PAL; Soares, Iriarte, et al., 2018; available at <http://p-pal.di.uminho.pt/tools>), three- to five-syllables are the most common EP words that, in addition, present more than 40 permissible first-syllable structures (see Soares, Iriarte, et al., 2018; see also Campos, Soares, & Oliveira, 2018; for further details). Together, these characteristics make EP a pivotal language to study not only the processes and mechanisms involved in reading acquisition as was the case hitherto (e.g. Fernandes et al., 2008; Seymour et al., 2003), but also the processes and mechanisms involved in skilled reading. However, to date, no studies have analyzed how these features impact the recognition and the pronunciation of EP words, by adopting a megastudy approach.

### The current study

The current study aimed to overcome this gap by using the megastudy approach and a series of regression analyses to explore the relative contribution that different orthographic, phonological and semantic variables play in the visual recognition and pronunciation of EP words. The collection of behavioural responses (reaction times and accuracy rates) from these tasks, considered the standard tasks to study processes of lexical access and word production, also allowed us to examine the similarities and the differences between the regressions models obtained from each task, as in previous

megastudies (e.g. Balota et al., 2004; Cortese & Khanna, 2007; Ferrand et al., 2011, 2017). In the regression analyses conducted, the orthographic, phonological and semantic variables shown to account for the highest percentages of variance in previous megastudies (e.g. Balota et al., 2004, 2007; Barca et al., 2002; Burani et al., 2007; Chateau & Jared, 2003; Cortese & Khanna, 2007; Cuetos & Barbón, 2006; Davies et al., 2013; Ferrand et al., 2010, 2011, 2017; González-Nosti et al., 2014; Keuleers et al., 2010, 2012; Wilson et al., 2013; Yap et al., 2010) were included as predictors.

Specifically, the standard measure of per million word frequency obtained both from EP written-texts (P-PAL database, Soares, Iriarte, et al., 2018) and EP subtitle corpora (SUBTLEX-PT database, Soares, Machado, et al., 2015) were used as orthographic predictors due to the relevance that word frequency plays in all research using verbal stimuli (see Brysbaert et al., 2011; and also Soares, Iriarte, et al., 2014; Soares, Machado, et al., 2015; for recent reviews). Note, that even though SUBTLEX-PT word counts were shown to represent a better determinant of EP lexical decision than P-PAL word counts (see Soares, Machado, et al., 2015), we decided to use both word frequency measures in the current study to further examine if the subtitle advantage observed by Soares, Machado, et al. (2015) in the visual word recognition of EP words is also observed in EP word pronunciation. Since subtitle word counts approach the day-to-day use of language more closely than written-texts, we expected the SUBTLEX-PT not only to outperform P-PAL word counts in naming performance as observed in other languages (e.g. Cai & Brysbaert, 2010; Cuetos et al., 2011), but also that the difference in the percentage of variance accounted for by each of these measures to be greater in naming than in lexical decision performance. Cuetos et al. (2011) and Cai and Brysbaert (2010) found that the advantage of subtitles over written-texts counts was greater in lexical decision than in naming performance in Spanish and Chinese, respectively. However, in these countries most films and TV series are dubbed and not subtitled. This is not the case of Portugal, where EP skilled readers are very used to read subtitles. Hence, this difference could lead to a different result in the EP language.

In addition, the Contextual Diversity (CD) measure obtained from the SUBTLEX-PT (Soares, Machado, et al., 2015), and the Zipf scale measure obtained both from the P-PAL (Soares, Iriarte, et al., 2018) and the SUBTLEX-PT (Soares, Machado, et al., 2015), were included in the analyses as orthographic predictors. CD is a more refined measure of word frequency that indexes the number of different contexts in which a word appears

and not simply the number of times a word appears regardless of the contexts of occurrence, as in the case of the standard per million word frequency (see Adelman, Brown, & Quesada, 2006; see also Soares, Machado, et al., 2015; for a discussion). The Zipf scale is another word frequency measure proposed recently by Van Heuven, Mandera, Keuleers, and Brysbaert (2014) that depicts word frequency in a 7-point logarithmic scale. Since it is a much easier and intuitive way to understand word frequency distribution (see Van Heuven et al., 2014), it has been increasingly used in experimental research (see Soares, Oliveira, Comesaña, & Costa, 2018; or Soares, Oliveira, Ferreira, et al., 2018; for recent examples). However, despite its usefulness, the empirical validation of the Zipf measure against other word frequency measures, namely the CD measure shown to be the best determinant of reading performance both in skilled and developing readers (e.g. Adelman et al., 2006; Brysbaert et al., 2011; Brysbaert & New, 2009; Keuleers et al., 2010; Perea, Soares, & Comesaña, 2013; Soares, Machado, et al., 2015; Van Heuven et al., 2014; see also Parmentier, Comesaña, & Soares, 2017; for evidence in serial recall tasks), has not been demonstrated.

Furthermore, since previous studies have shown that the phonetic features of words' accounted for a significant percentage of variance in speeded pronunciation (e.g. Balota et al., 2004; Chateau & Jared, 2003; Cortese & Khanna, 2007; Davies et al., 2013; Ferrand et al., 2011; Spieler & Balota, 2000; Treiman et al., 1995; Yap & Balota, 2009), we examined the role that words' first-phoneme and words' stress pattern play in EP word processing. Note that although the abovementioned studies showed that words' first-phoneme and words' stress pattern impact strongly word pronunciation than word visual recognition, the fact that in EP the vast amount of orthography-phonology inconsistencies are contextually resolved (for instance the letter "s" at first position as in *selo*[stamp] is always pronounced [s], while when it appears at the middle position between vowels, as in *casa*[house], it is always pronounced [z]) might affect lexical decision in a greater extent than observed in these studies, particularly those conducted in deep languages (e.g. Balota et al., 2004; Chateau & Jared, 2003; Cortese & Khanna, 2007; Spieler & Balota, 2000; Treiman et al., 1995; Yap & Balota, 2009).

Likewise, word length, measured in number of letters, phonemes, and syllables (orthographic and phonological) were also added to the analyses as predictors. Even though reliable word length effects were not consistently observed across-languages and tasks (see New et al., 2006; for a review), the length of a printed word is a critical parameter in most models of visual word recognition (see Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001;

Frost, 1998; Goswami & Ziegler, 2006; Perfetti et al., 1992; Ziegler & Goswami, 2005). Larger word length effects have been observed in naming than in lexical decision performance and also in languages with shallow than deep orthographies (e.g. Baayen et al., 2006; Balota et al., 2004; Barca et al., 2002; Burani et al., 2007; Chateau & Jared, 2003; Cortese & Khanna, 2007; Cuetos & Barbón, 2006; Davies et al., 2013; Ferrand et al., 2011; González-Nosti et al., 2014; Spieler & Balota, 2000; Wilson et al., 2013; Yap et al., 2010; Yap & Balota, 2009). These results have been interpreted as a marker of the level of engagement of the sub-lexical route in word processing. Indeed, since the use of this route requires the letter string to be segmented into their basic components (e.g. letters), and then converted into sounds by using a set of grapheme-phoneme correspondence rules, it is expected that the time needed to process a word would increase as a function of the number of letters in the string, particularly in shallow orthographies as Spanish or Italian. Nevertheless, word length effects have been also observed in languages with lower levels of orthography-phonology consistency such as English or French, and not only in word pronunciation but also in word/nonword decisions (e.g. Baayen et al., 2006; Balota et al., 2004; Chateau & Jared, 2003; Cortese & Khanna, 2007; Ferrand et al., 2011; New et al., 2006; Spieler & Balota, 2000; Wilson et al., 2013; Yap et al., 2010; Yap & Balota, 2009).

These inconsistencies can be explained, at least in part, by the fact that the relationship between word length and word latencies is not linearly defined, as demonstrated by New et al. (2006) with the lexical decision data from the English Lexicon Project. Specifically, the authors found a U-shaped function that was independent of word frequency, number of syllables, and number of orthographic neighbours measured as the number of words that can be formed by replacing a given letter in the string by another letter, while keeping the remaining constant in the same positions, i.e. the standard  $N$  orthographic neighbourhood measure of Coltheart, Davelaar, Jonasson, and Besner (1977). Facilitative length effects were observed for words from three- to five-letters, null effects for words from five- to eight-letters, and inhibitory length effects for words from eight- to-13-letters. This U-shaped function was also observed by Ferrand et al. in the lexical decision data from both the French Lexicon Project (Ferrand et al., 2010), and the recent MEGALEX database (Ferrand et al., 2017; see however Ferrand et al., 2011; for a J-shaped function in the lexical decision data from Chronolex), although, in the French language, the effect vanished when the most recent OLD<sub>20</sub> orthographic neighbourhood measure (Yarkoni et al., 2008)

was partialled out. The  $OLD_{20}$  is a richer and more flexible way of measuring orthographic similarity, as it indexes the mean number of operations necessary to transform a word into another word in the lexicon considering the 20 closest orthographic neighbours. As such, it avoids the negative relationship that between the number of letters and the number of orthographic neighbours is observed when the classic  $N$  metric is used instead (see Yarkoni et al., 2008; for details).

Additionally, it is also worth noting that inhibitory word length effects were also observed in the English language when larger word length units were considered (number of syllables), though the effect was, in this case, linearly defined (see New et al., 2006; see also Yap & Balota, 2009; for further evidence). These results are consistent with the assertions of the grain-size theory (Goswami & Ziegler, 2006; Ziegler & Goswami, 2005), claiming that in nonshallow orthographies, phonological recoding involves the use of multiple and larger size units (e.g. syllables) as in these languages the use of small units (i.e. letters) is more prone to error than in shallow orthographies. Therefore, in EP, we expected to observe inhibitory length effects both in lexical decision and naming performance. Note, however, that in a previous factorial study conducted with EP skilled readers, Lima and Castro (2010) only found reliable word length effects in lexical decision for EP words differing in two letters (i.e. words with four- vs. words with six-letters) and in naming performance only when words and nonwords were presented in mixed lists (i.e. in a context that stimulates the use of the serial sub-lexical route of processing). Nonetheless, since that study only used a limited set of EP words from a very narrow word length range (i.e. 100 dissyllabic words from four- to six-letters), it is important to further examine whether these results can be also observed when a more diversified set of EP words coming from a wide length range were used. Moreover, it is also critical to analyze whether EP word length effects would be also observed when larger (syllables) units are analyzed. Indeed, although the characteristics of the EP language made syllabic boundaries to be less clearly defined in EP than in other languages as mentioned, recent EP factorial studies have also demonstrated that the syllable plays a functional role both in visual recognition and pronunciation of EP words (e.g. Campos, Oliveira, et al., 2018; Campos, Soares, et al., 2018; Pureza, Soares, & Comesaña, 2016), thereby suggesting that EP skilled readers may rely on the use of multiple recoding strategies, as English skilled readers (e.g. Goswami & Ziegler, 2006; New et al., 2006; Yap & Balota, 2009; Ziegler & Goswami, 2005).

Furthermore, it remains also unclear whether EP word length effects would survive when word neighbourhood measures (e.g.  $N$  metric,  $OLD_{20}$ ) were taken into account, as in New et al. (2006) and Ferrand et al. (2010, 2011, 2017) studies. As advanced by Ferrand et al. (2010), since shorter words present more neighbours than longer words, it is possible that the word length effects observed in different languages may result from the number of neighbours in the lexicon and not from word length per se. Although the magnitude and the direction of neighbourhood effects seem to depend on the type of neighbourhood measure considered, on the characteristics of the language in use, as well as on task demands, facilitative neighbourhood effects tend to be observed both in lexical decision, and particularly in naming performance in different languages (see Andrews, 1997; Perea, 2015; for reviews). These results have been taken as evidence that words from denser orthographic neighbourhoods present more stable lexical representations than words from sparse neighbourhoods, and also that retrieving phonological information from an orthographic input (as in speed pronunciation) relies more on the use of the sub-lexical of processing, which seems to be more sensitive to the orthographic and/or phonological similarities among words in the lexicon than word/nonword discriminations. Nonetheless, larger neighbourhood effects have been observed on the  $OLD_{20}$  than on the  $N$  measure (e.g. Yarkoni et al., 2008; Yap & Balota, 2009; see however Ferrand et al., 2010, 2017; or Keuleers et al., 2010; for a modest contribution of  $OLD_{20}$  in the lexical decision data from French and British-English participants, and also Yap et al., 2010 for similar results in the speeded pronunciation data from Malay participants), and also in studies with factorial than megastudy designs. As pointed out by Keuleers et al. (2012), this might arise because in factorial studies researchers tend to use words with extreme neighbourhood values (i.e. words with a very few or a very high number of orthographic and/or phonological neighbours), which increases the likelihood of observing strong neighbourhood effects.

Additionally, recent megastudies (e.g. Ernestus & Cutler, 2015; Ferrand et al., 2017; Goh et al., 2016) have also analyzed the role that uniqueness point measures (i.e. the point [letter/phoneme] in the string from which a word becomes distinguishable from all its neighbours – e.g. Kwantes & Mewhort, 1999; Luce, 1986) play in word recognition in different languages. Classic factorial studies showed that words with an early orthographic or an early phonological uniqueness point tend to be recognised and pronounced faster and more accurately than words with a late orthographic or a late

phonological uniqueness point (see Kwantes & Mewhort, 1999; Luce, 1986; or Radeau, Mousty, & Bertelson, 1989; for some examples). This effect has been associated with the use of the serial left-to-right route of processing (see however Izura, Wright, & Fouquet, 2014; for a factorial study showing an inhibitory orthographic uniqueness point effect in the English language). Yet, the results from the few megastudies testing these measures are inconclusive. Using different phonological uniqueness point measures, Ernestus and Cutler (2015) found that word duration was the best predictor of the latencies with which Dutch spoken words were recognised, and Goh et al. (2016) observed that the phonological uniqueness point measure only contributed (marginally) for the speed with which English spoken words were semantically categorised. Still, Ferrand et al. (2017), showed that the effect of the orthographic and phonological uniqueness point measures seems to depend on the modality (visual vs. auditory) in which French words were presented. Using the visual- and the auditory-lexical decision data from the MEGALEX database, the authors demonstrated that, in the visual domain, words with an early orthographic or phonological uniqueness point were recognised more slowly than words with a late orthographic or phonological uniqueness point (an inhibitory effect consistent with the results of Izura et al., 2014). However, in the auditory domain, words with an early phonological uniqueness point produced faster recognition times, whereas words with an early orthographic uniqueness point produced longer recognition times. Thus, the effects of the orthographic and phonological uniqueness point measures in shallow and deep orthographies are unclear, and completely unknown in intermediate-depth languages, as EP.

Finally, since semantic variables as imageability (i.e. the ease and speed with which a word evokes a mental image – e.g. Paivio, Yuille, & Madigan, 1968), concreteness (i.e. the degree to which words refer to objects, persons, places, or things that can be experienced by the senses – e.g. Paivio et al., 1968), subjective frequency (i.e. the estimated number of times a word is encountered in its spoken or written form by individuals in their daily lives, Balota, Pilotti, & Cortese, 2001), AoA (i.e. the estimated age at which a word is learned by individuals, Carroll & White, 1973), and words' affective properties such as valence (i.e. the degree of pleasantness a word evokes in individuals), arousal (i.e. the degree of physiological activation it triggers) and dominance (i.e. the degree of control it produces) (see Bradley & Lang, 1994; see also Soares, Comesaña, Pinheiro, Simões, & Frade, 2012; or Pinheiro, Dias, Pedrosa, & Soares, 2017; for recent studies collecting these emotional ratings for EP verbal stimuli) were shown to account for significant,

though often small (around 1%–5%), amounts of variance in word processing (e.g. Balota et al., 2004; Bonin et al., 2018; Cortese & Khanna, 2007; Cortese & Schock, 2013; Davies et al., 2013; Ferrand et al., 2011, 2017; Goh et al., 2016; González-Nosti et al., 2014; Kousta et al., 2011; Kuperman, 2015; Kuperman et al., 2012; Wilson et al., 2013; Yap & Balota, 2009), they were also introduced in the analyses as semantic predictors. Note that although subjective and objective word frequency measures correlates strongly, subjective frequency was shown to be a better determinant of word recognition than objective word counts (e.g. Balota et al., 2001, 2004; Cortese & Khanna, 2007). Moreover, subjective frequency was also shown to be a better predictor of the relative frequency of exposure to a word in daily life than the experiential familiarity construct introduced by Gernsbacher (1984) (see Balota et al., 2001), hence making the inclusion of this variable relevant for the purposes of the current paper. Furthermore, it is also worth noting that despite the lively debate concerning the role that AoA assumes in word processing and how AoA should be measured (see Ghyselinck, Lewis, & Brysbaert, 2004; Zevin & Seidenberg, 2002; or Soares, Medeiros, et al., 2014; for a discussion), here we opted to include this variable under the “semantic” category as in other works (e.g. Kuperman et al., 2012; Soares, Costa, Machado, Comesaña, & Oliveira, 2017) because AoA correlates more strongly with semantic variables such as imageability or concreteness than with lexical variables such as objective word frequency (e.g. Brysbaert, Van Wijnendaele, & De Deyne, 2000; Cortese & Khanna, 2007; Steyvers & Tenenbaum, 2005).

Overall, evidence from studies testing the role of these variables in word processing demonstrate that words that are acquired earlier in life, that are more concrete (see, however, Bonin et al., 2018; Kousta et al., 2011; for a concreteness reverse effect), more imaginable, more familiar, and more positively valenced, are named and recognised faster and more accurately than words presenting lower scores on these dimensions (see Soares et al., 2017; for a recent review). Stronger semantic effects have been also observed in lexical decision than in naming performance (e.g. Balota et al., 2004; Bonin et al., 2018; Cortese & Khanna, 2007; Cortese & Schock, 2013; Ferrand et al., 2011, 2017; Wilson et al., 2013; Yap & Balota, 2009), which has been taken as reflecting a stronger reliance on the meaningfulness of the stimulus in word/nonword discriminations than in speed pronunciation. Indeed, since words' phonological information can be successfully achieved through the use of grapheme-phoneme conversion rules, particularly in languages with shallow orthographies, naming performance is expected to be less affected by semantic-

to-orthographic/phonological feedback connections than word/nonword discriminations (see Balota, Ferraro, & Connor, 1991; for details). Note, however, that semantic effects have been also observed in naming studies and not only from deep but also from shallow orthographies (e.g. Balota et al., 2004; Cortese & Khanna, 2007; Cuetos & Barbón, 2006; Davies et al., 2013; Ferrand et al., 2011; Goh et al., 2016; Wilson et al., 2013; Yap et al., 2011; Yap & Balota, 2009). In intermediate-depth orthographies, studies aiming to analyze how semantic variables affect word processing are, to the best of our knowledge, inexistent. The absence of semantic norms for EP word stimuli certainly contributed to this situation. However, norms for imageability, concreteness and subjective frequency, as well for AoA and for the affective dimensions of valence, arousal and dominance, have become recently available for EP from the Minho Word Pool database (Soares et al., 2017), the Cameirão and Vicente (2010) norms, and the Portuguese adaptation of the Affective Norms for English Words (Bradley & Lang, 1999; Soares et al., 2012). Therefore, these ratings were used in the current paper to explore how these variables affect visual recognition and pronunciation of EP words. As in other languages, we expected semantic variables to account for significant (though small) amounts of variance of EP skilled readers' performance, particularly in word/nonword discriminations, after all the other variables shown to affect EP word processing were controlled for.

## Method

### Participants

A total of 110 Portuguese college students (96 females), with ages between 18 and 32 years ( $M = 21.0$ ,  $SD = 3.32$ ) participated in the study. All of them were native speakers of EP and reported normal audition and normal or corrected-to-normal visual acuity, as well as no history of language or learning disabilities. The majority was right-handed (92%). Half of the participants ( $M_{age} = 21.2$ ,  $SD = 2.87$ , 49 females) performed the visual lexical decision task (LDT), while the other half ( $M_{age} = 20.8$ ,  $SD = 3.73$ , 47 females) performed the naming task (NAM). Participants were randomly assigned to each task while ensuring the same number of participants per task ( $n = 55$ ). Participants were informed that the participation in the experiment (either LDT or NAM) involved the collection of data in four consecutive sessions, each separated by a week. Only participants who completed the four sessions received course credits for their participation. The experiment was conducted with the approval of the Ethics Committee of the University

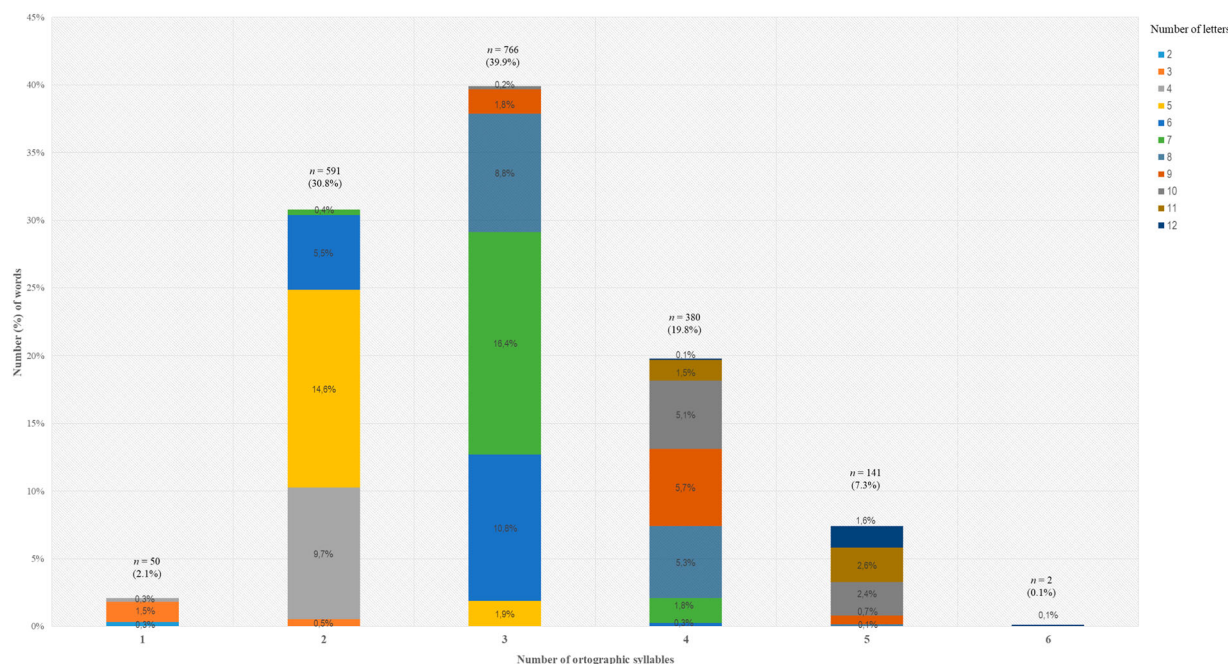
of Minho. Written informed consent was obtained from all the participants.

### Materials

The stimuli selected for both experiments consisted of 1,920 EP words plus 1,920 pronounceable EP nonwords in the case of the LDT. Words were selected from the Minho Word Pool database (Soares et al., 2017). Nonwords were created by changing one-to-five letters in the non-terminal positions from other EP words matched with the 1,920 experimental words in word length (in number of letters and syllables) and word frequency based on the P-PAL database (Soares, Iriarte, et al., 2018). For instance, the nonword *catufia* was created from the base-word *calúnia*[slander], respecting the phonotactic restrictions of the Portuguese language. This method followed many other studies conducted in EP (e.g. Campos, Oliveira, et al., 2018; Campos, Soares, et al., 2018; Perea et al., 2013; Perea, Comesaña, & Soares, 2012; Soares, Lages, Oliveira, & Hernández, 2018; Soares, Machado, et al., 2015; Soares, Perea, & Comesaña, 2014) and in many other languages (e.g. Balota et al., 2004, 2007; Cuetos et al., 2011; Ferrand et al., 2010, 2011; Sze et al., 2014; Yap et al., 2010), as research tools supporting the generation of legal nonwords are not available for EP. Nevertheless, extreme caution was taken to ensure that the nonwords resembled real EP words as much as possible.

The 1,920 words selected were intentionally chosen to closely represent the diversity of the EP language. Indeed, although several large-scale studies used only mono- and disyllabic words as stimuli (e.g. Baayen et al., 2006; Balota et al., 2004; Cortese & Khanna, 2007; Goh et al., 2016; Keuleers et al., 2010, 2012; Sze et al., 2014; Tse et al., 2017), most of the EP words included in both experiments were multisyllabic, as most EP words present more than two syllables (see Soares, Iriarte, et al., 2018). Figure 1 presents the distribution of the EP words selected in the dataset as a function of the number of syllables and number of letters within each syllable length.

As shown in Figure 1, monosyllables corresponded only to 2.1% of the words in the dataset. Most words had two to three syllables ( $n = 1,357$ , 70.7%). Words with more than three syllables corresponded to 27.2% of the total number of words ( $n = 523$ , 27.2%). Concerning the number of letters, words from five to eight letters were the most common ( $n = 1,264$ , 65.8%), followed by words from nine to 12 letters ( $n = 419$ , 21.8%), and, finally, by words presenting two to four letters ( $n = 237$ , 12.3%). Moreover, an equivalent number of low-frequency (i.e. <10 occurrences per



**Figure 1.** Distribution of the EP words as a function of number of syllables and letters. Both number of letters and number of orthographic syllables were obtained from the P-PAL database (Soares, Iriarte, et al., 2018).

million words), medium-frequency (i.e. 11–74 occurrences per million words) and high-frequency words (i.e.  $\geq 75$  occurrences per million words), based on the P-PAL database (Soares, Iriarte, et al., 2018), were also included for short (two to four letters and one to two syllables), medium (five to eight letters and two to four syllables) and long (nine to 12 letters and three to six syllables) EP words,  $\chi^2(4) = 5.79$ ,  $p = .22$ . Therefore, although this dataset integrates a smaller number of words than observed in other large-scale studies (particularly the lexicon projects mentioned before), the fact that these words were intentionally selected to represent the lexical diversity in the EP language makes this dataset a powerful tool to generalise the results obtained here to the other words in the EP lexicon.

Table 1 presents the descriptive statistics for the orthographic, phonological and semantic measures that were included in the multiple regression analyses conducted in this paper and presented in the next sections. The distribution of the different measures under these categories (i.e. orthographic, phonological, semantic) followed the options adopted by other authors in previous large-scale studies (e.g. Balota et al., 2004, 2007; Barca et al., 2002; Brysbaert et al., 2016; Burani et al., 2007; Cortese & Khanna, 2007; Cuetos et al., 2011; Davies et al., 2013; Ernestus & Cutler, 2015; Ferrand et al., 2010, 2011, 2017; Goh et al., 2016; González-Nosti et al., 2014; Keuleers et al., 2010, 2012; Sze et al., 2014, 2015; Tsang et al., 2017; Tse et al., 2017; Yap et al.,

2010). Values are provided in Table 1 for the total number of words for which behavioural responses (reaction times and accuracy) were collected ( $N = 1,920$ ), except for AoA, valence, arousal and dominance affective dimensions because, for these latter variables, EP norms are only available for 818 (AoA) and 484 (valence, arousal, and dominance) words from our dataset. Written-text word frequency, orthographic and phonological statistics were obtained from the P-PAL database (Soares, Iriarte, et al., 2018). Subtitle word frequency measures were obtained from the SUBTLEX-PT database (Soares, Machado, et al., 2015). The phonological characteristics of the EP words (onsets and stress pattern) were categorised dichotomously, in which “1” denotes the presence of a given phonetic feature and “0” the absence of that feature, following the procedures adopted in other studies (e.g. Balota et al., 2004; Chateau & Jared, 2003; Cortese & Khanna, 2007; Davies et al., 2013). Seven categories were used to classify the phonetic features of the EP words beginning by a vocalic phoneme (front, central, back, high, mid, low, and rounding) – corresponding to  $\cong 23\%$  of the words; and 11 to classify the words beginning by a consonant phoneme (bilabial, labiodental, apico-dental, alveolar, palatal, velar, stop oral, stop nasal, fricative, lateral and voiced). The stress pattern was classified as paroxytone (words stressed on the penultimate syllable), oxytone (words stressed on the ultimate syllable) and propoxytone (words stressed on the antepenultimate syllable).



**Table 1.** Psycholinguistic characteristics of the words selected for the lexical decision (LDT) and naming (NAM) tasks on the orthographic, phonological and semantic variables included in the by-item regression analyses.

Word measures		M	SD	Range (min.-max.)	
Orthographic measures	P-PAL <sub>pmwff</sub>	67.31	110.69	.43–1,214.45	
	P-PAL <sub>Zipf</sub>	4.46	.58	2.64–6.08	
	SUBTLEX-PT <sub>pmwff</sub>	59.77	138.49	.06–1,789.43	
	SUBTLEX-PT <sub>Zipf</sub>	4.30	.64	1.81–6.26	
	SUBTLEX-PT <sub>CD</sub>	.07	.08	.00–.51	
	P-PAL <sub>N<sub>lett</sub></sub>	6.90	2.10	2–12	
	P-PAL <sub>NO<sub>syll</sub></sub>	3.00	.95	1–6	
	P-PAL <sub>OLD<sub>20</sub></sub>	1.94	.57	1.00–3.85	
	P-PAL <sub>ON</sub>	3.73	4.79	0–27	
	P-PAL <sub>OUP</sub>	6.84	2.06	2–12	
Phonological measures	P-PAL <sub>N<sub>phon</sub></sub>	6.48	1.96	2–12	
	P-PAL <sub>NP<sub>syll</sub></sub>	2.92	.90	1–6	
	P-PAL <sub>PN</sub>	4.20	5.10	0–30	
	First-phoneme characteristics (onsets)	Front	.07	.26	0–1
		Central	.13	.34	0–1
		Back	.03	.17	0–1
		High	.10	.30	0–1
		Mid	.12	.32	0–1
		Low	.02	.14	0–1
		Rounded	.03	.17	0–1
		Bilabial	.21	.41	0–1
		Labiodental	.09	.28	0–1
		Apico-dental	.18	.39	0–1
		Alveolar	.06	.23	0–1
		Palatal	.02	.15	0–1
		Velar	.21	.40	0–1
		Stop oral	.41	.49	0–1
		Stop nasal	.08	.27	0–1
		Fricative	.19	.39	0–1
		Lateral	.09	.28	0–1
Voiced	.33	.47	0–1		
Word stress pattern	Paroxytone	.73	.445	0–1	
	Oxytone	.21	.410	0–1	
	Proparoxytone	.06	.233	0–1	
Semantic measures	Imag (1–7 scale)	4.52	1.05	2.29–6.75	
	Conc (1–7 scale)	4.51	1.27	1.87–6.91	
	SubjFreq (1–7 scale)	4.74	.90	1.68–6.93	
	AoA (1–9 scale)	5.26	1.56	1.33–8.34	
	Val (1–9 scale)	5.43	1.78	1.34–8.38	
	Arou (1–9 scale)	4.78	1.13	1.79–7.65	
	Dom (1–9 scale)	5.16	.88	1.95–7.47	

Notes: P-PAL<sub>pmwff</sub>: per million word frequency, and P-PAL<sub>Zipf</sub>: Zipf scale word frequency obtained from the P-PAL database (Soares, Iriarte, et al., 2018); SUBTLEX-PT<sub>pmwff</sub>: per million word frequency, SUBTLEX-PT<sub>Zipf</sub>: Zipf scale word frequency, and SUBTLEX-PT<sub>CD</sub>: Contextual Diversity word frequency as obtained from the SUBTLEX-PT database (Soares, Machado, et al., 2015);  $N_{lett}$ : Number of letters,  $NO_{syll}$ : Number of orthographic syllables,  $OLD_{20}$ : Orthographic Levenshtein Distance,  $ON$ : Orthographic Neighbourhood size,  $OUP$ : Orthographic Uniqueness Point,  $N_{phon}$ : Number of phonemes,  $NP_{syll}$ : Number of phonological syllables and  $PN$ : Phonological Neighbourhood size as obtained from the P-PAL database (Soares, Iriarte, et al., 2018). Imag: Imageability; Conc: Concreteness; and SubjFreq: Subjective frequency obtained from the Minho Word Pool database (Soares et al., 2017); Val: Valence; Arou: Arousal; and Dom: Dominance obtained from Soares et al. (2012) norms; AoA: Age of Acquisition obtained from Cameirão and Vicente (2010) norms.

## Procedure

### LDT

Each participant responded to 3,800 stimuli (1,920 words and 1,920 nonwords) in four experimental sessions that took place in four consecutive weeks (one session per week), with a quarter of the stimuli presented in each session (four blocks with 960 stimuli – 480 words and 480 nonwords – each). Stimuli presented in each block were randomly selected from the total stimulus pool, with the constraint that a similar number of words from different lengths (short, medium, and long words) and frequency intervals (low, medium, high) were included in each block (Block 1:  $\chi^2(4) = 3.24$ ,  $p = .52$ ; Block 2:  $\chi^2(4) = 4.60$ ,  $p = .33$ ; Block 3:  $\chi^2(4) = 3.76$ ,  $p = .44$ ;

and Block 4:  $\chi^2(4) = .38$ ,  $p = .98$ ). Nonwords were assigned to each block according to the characteristics (word length and word frequency) of the basewords used to generate the nonwords. Block presentation was counter-balanced across participants (24 different orders). Participants were randomly assigned to each order (approximately two participants per order). The experiment was run individually in a soundproof booth in each of the four experimental sessions at the facilities of the Human Cognition Laboratory (University of Minho). Stimulus presentation and response recording (reaction times and accuracy) were controlled with DMDX software (Forster & Forster, 2003). In each experimental session, participants were asked to decide as

quickly and accurately as possible if the string of letters presented at the centre of the computer screen in lowercase (Courier New 14) was or was not a real EP word, by pressing the “Z” key on the keyboard to a *sim*[yes] response, and the “M” key on the keyboard to a *não* [no] response. Both speed and accuracy were stressed in the instructions. In each block, trials were randomly presented. Each trial consisted of a sequence of three visual events. The first was a fixation point (+) presented at the centre of the computer screen for 500 ms. The fixation point was immediately replaced by the stimulus (word or nonword) at the same position and remained on the screen until a participant’s response or until 2,500 ms had elapsed. The next trial began after an inter-trial interval (blank screen) of 500 ms. Participants were also informed about the existence of pauses after every 80 trials. Prior to the presentation of the trials, participants received 12 practice trials (6 words and 6 nonwords) to familiarise them with the task (different across blocks). Each experimental session lasted approximately 45 min. The entire procedure lasted about 3 h per participant.

### NAM

As in the LDT task, each participant responded to the stimuli (words) in four consecutive experimental sessions separated by a week each (one block per session). The blocks used were the same of the LDT, with the exception that nonwords were not presented in this task. Thus, in each session, participants responded to 480 words (1,920 in total). Blocks were counterbalanced across participants following the same procedure as in the LDT (approximately two participants per order). Data were collected individually in soundproof booths at the facilities of the Human Cognition Laboratory (University of Minho) in each of the four experimental sessions. Stimulus presentation was controlled with the DMDX software (Forster & Forster, 2003). Naming latencies were recorded with voicekey from the presentation of the word to the onset of the naming response. Accuracy and response times of the recorded vocal responses were checked offline by using the CheckVocal software (Protopapas, 2007). Participants were instructed to read out loud as quickly and as accurately as possible the words that were presented at the centre of the computer screen. They were also instructed to avoid producing any extraneous noises that could trigger the voicekey. Trials in each block were randomly presented. Each trial consisted of the following sequence of events: a fixation point (+) presented at the centre of the computer screen in lowercase (Courier New 14) for 500 ms, the word to be named at the same position for 2,000 ms or until the voicekey registered an acoustic signal, and a blank screen for 500 ms that

worked as inter-trial interval. In each block, participants were informed about the existence of pauses after every 80 trials. Prior to the experiment, six practice trials were used to familiarise participants with the task (different across blocks). Each experimental session lasted approximately 30 min. The entire procedure lasted about 2 h per participant.

## Results and discussion

### Trimming procedures

Before computing the lexical decision and the naming data (latency and accuracy) for the correct responses of each of the 1,920 EP words used in this megastudy, several trimming procedures were implemented on the raw data that comprised 211,200 responses in the LDT data and 105,600 responses in the NAM data. Participants showed a high accuracy rate in both tasks (96.86% in LDT and 99.43% in NAM) for word trials. Thus, all participants ( $N = 55$ ) were included in the computation of mean reaction times (RT) and accuracy (%Acc) rates presented in Table 2 and in the excel file that can be downloaded as a supplemental archive from this paper or at <http://p-pal.di.uminho.pt/about/databases>. Following a common practice in megastudies (e.g. Balota et al., 2007; Ferrand et al., 2010, 2017), RTs that were shorter than 200 ms or longer than 2,000 ms (1,500 in the naming data) were excluded from the latency analyses, as well as those that were 2.5 SDs above or below the mean RTs of each participant. This procedure resulted in the exclusion of 2.70% responses in the LDT latency data and 2.82% in the NAM latency data. Therefore, the RT values provided were based on 99,429 valid responses in LDT data (corresponding to 94.16% of the original word data), and on 102,016 valid responses in NAM data (corresponding to 96.61% of the original data).

### Reaction times and accuracy

Table 2 presents the descriptive statistics for the raw RTs (in ms), as well as for the z-transformations of the raw RTs

**Table 2.** Descriptive statistics of the latency (RT, zRT) and accuracy (%Acc) data obtained for the correct word responses from the lexical decision (LDT) and naming (NAM) tasks.

Task	Behavioural measures	Min	Max.	Mean	SD
LDT	Raw latencies (RTs)	485.70	748.32	561.25	39.36
	Standardised latencies (zRTs)	−.69	1.94	.02	.39
	Accuracy (%Acc)	16.40	100.00	96.86	5.24
	Number of observations	4	55	51.79	3.91
NAM	Raw latencies (RTs)	464.21	657.81	558.06	31.58
	Standardised latencies (zRTs)	−1.20	1.52	.01	.43
	Accuracy (%Acc)	80.40	100.00	99.43	1.21
	Number of observations	19	55	52.24	2.47

obtained for each of the 1,920 words included in this dataset (item statistics) both for LDT and NAM data. Z-scores were calculated based on the standardised latencies of the raw scores obtained by each participant (see Balota et al., 2007; Ferrand et al., 2010, 2011, 2017; or Keuleers et al., 2010, 2012; for similar procedures) in each of the tasks. Z-scores allow a direct comparison of the RTs obtained for each word and, in addition, minimise variability in item performance due to individual differences (see Faust, Balota, Spieler, & Ferraro, 1999). Moreover, for each task, accuracy, i.e. the percentage of correct responses (%Acc) and the number of observations per item are also presented.

The results depicted in Table 2 show that the behavioural data obtained in the current study are in accordance with other chronometric datasets available for other languages (e.g. Balota et al., 2004, 2007; Davies et al., 2013; Ferrand et al., 2010, 2011, 2017; González-Nosti et al., 2014). Nevertheless, it is worth noting that EP participants showed faster RTs and higher accuracy rates than those reported in previous large-scale studies, particularly in the LDT (e.g. Balota et al., 2007; Ferrand et al., 2010, 2011; Keuleers et al., 2010, 2012). Differences in the items used, in data collection procedures and participants' recruitment strategy might account for these discrepancies. Note, for instance, that the words used in the current study were, on average, more frequent and shorter than those used in the ELP or in the FLP (see Balota et al., 2007; Ferrand et al., 2010; for details). Furthermore, the nonwords were generated by changing as many letters as the length of the corresponding basewords, thus maintaining nonwordlikeness constant across words' length (see Ferrand et al., 2010 for a similar procedure). Moreover, our data were collected in a laboratory setting under very strict control conditions, and only college students from the same university were included in the experiments. Finally, the fact that in other large-scale studies (e.g. Ferrand et al., 2017; Keuleers et al., 2010, 2012) participants were informed that if their accuracy fell below 85% their payment would suffer penalties may have also contributed to make participants more cautious about their responses, and, hence, to present longer latencies. Note, however, that although EP participants presented shorter latencies than participants from other languages, the speed with which EP words were recognised and pronounced did not compromise the accuracy with which EP words were recognised/pronounced as they were above those observed in previous studies (e.g. Balota et al., 2007; Ferrand et al., 2010, 2011; Keuleers et al., 2010, 2012). Nevertheless, the inspection of the accuracy rates per item showed that nine words in the dataset were incorrectly recognised by at least one

third of the participants in the LDT, including words as *cárcere*[prison cell], *asilo*[asylum], *lodo*[sludge] and *zelo* [zeal], all presenting a very low lexical frequency both in the P-PAL (Soares, Iriarte, et al., 2018) and SUBTLEX-PT (Soares, Machado, et al., 2015) databases. Therefore, following a common practice in megastudies (e.g. Balota et al., 2004; Ferrand et al., 2010, 2017), we excluded these words from the regression analyses conducted both on the LDT and NAM data. Even though all words in NAM presented an accuracy level above 67%, we opted to exclude these nine words from the NAM data to make the results from both tasks fully comparable. Excluding these words from the dataset increased the accuracy rates to 97.07% in LDT data and to 99.45% in the NAM data, and the mean number of observations *per* item to 51.92 ( $SD = 3.35$ , min. = 26, max. = 55) in the TDL data, and to 52.28 ( $SD = 2.32$ , min. = 31, max. = 55) in the NAM data.

### Reliability

Reliability analyses were conducted using both the split-half method based on the odd and the even groups of participants that performed each task (see Ferrand et al., 2010; or Keuleers et al., 2010 for a similar procedure), and on the intra-class correlation coefficients (ICCs) proposed by Rey and Courrieu (2010) and recently used by Keuleers et al. (2012). This seems to be the best procedure to deal with missing data from megastudies, particularly those using a between-subjects design (i.e. in which different participants respond to a different pool of items) as in the British Lexicon Project (Keuleers et al., 2010). Results from the two methods showed that the behavioural data were highly reliable in both tasks. Specifically, the split-half method with the Spearman-Brown correction for length showed that the reliability of the LDT data was .869 for the raw RTs, .887 for the zRTs and .829 for %Acc. In NAM, reliability scores were even higher in the latency data (.898 for raw RTs, and .919 for zRTs), though lower in accuracy (.341). This result indicates that the low variability in the accuracy NAM data could be replicated, probably due to the ceiling effects observed in the NAM performance. Results from the ICCs method also showed highly satisfactory reliability scores, mimicking the results obtained with the split-half method (LDT: .869 for raw RTs, .885 for zRTs and .827 for %Acc; NAM: .898 for raw RTs, .916 for zRTs and .341 for %Acc). Note that since the missing data were low in both tasks, it is not surprising that the results from both methods were virtually the same. As in previous megastudies (e.g. Balota et al., 2007; Ferrand et al., 2010; Keuleers et al., 2010, 2012; Tse et al., 2017), zRTs produced better reliability results than raw

RTs in both tasks, although the gain from removing individual differences in the overall RT data was less pronounced here than in the abovementioned studies. The fact that a within-subjects design was used (i.e. all participants responded to the whole pool of stimuli), along with the use of a more homogeneous sample of skilled readers (see for instance Balota et al., 2007; or Keuleers et al., 2010, 2012), might have contributed to attenuate differences in item performance across participants. Taken together, these results provide strong support for the item estimates provided in the EP chronometric dataset both for LDT and NAM performance and suggest that the analyses conducted based on this chronometric dataset are highly reliable.

### Practice effects

Since data were collected in different experimental sessions, we explored practice effects in participants' performance (RTs) as in previous megastudies (e.g. Ferrand et al., 2017; Keuleers et al., 2010, 2012). For that purpose, the speed with which words were correctly responded in each block was examined, by considering the order by which they were responded across sessions by each participant. As observed in other megastudies (e.g. Ferrand et al., 2017; Keuleers et al., 2010, 2012), practice effects were observed both on LDT,  $F(3, 162) = 21.519$ ,  $MSE = 625.09$ ,  $p < .001$ ,  $\eta_p^2 = .29$ , and NAM performance,  $F(3, 162) = 4.15$ ,  $MSE = 633.82$ ,  $p = .007$ ,  $\eta_p^2 = .07$ . In LDT, the effect showed that participants were approximately 30 ms slower responding to words from the first block than to words from any other block ( $p < .001$ ), with the differences between the other blocks being statistically non-significant ( $M_{\text{Block1}} = 583.03$ ;  $M_{\text{Block2}} = 549.99$ ;  $M_{\text{Block3}} = 552.49$ ;  $M_{\text{Block4}} = 553.12$ ). In NAM, the effect showed that participants were approximately 16 ms slower in responding to words from the first than to words from the fourth block ( $p = .033$ ), being the remaining comparisons non-significant ( $M_{\text{Block1}} = 550.96$ ;  $M_{\text{Block2}} = 556.21$ ;  $M_{\text{Block3}} = 562.07$ ;  $M_{\text{Block4}} = 566.86$ ). However, when the counterbalancing design used in data collection was considered, practice effects vanished both from the LDT,  $F(3, 162) = .137$ ,  $MSE = 871.98$ ,  $p = .938$ ,  $\eta_p^2 = .003$ , and NAM data,  $F(3, 162) = .333$ ,  $MSE = 678.38$ ,  $p = .802$ ,  $\eta_p^2 = .006$ . Thus, the use of a counterbalanced Latin-Square design seems to have been effective in removing variability from the RT data due to participants' familiarisation with the task across sessions – note, however, that since Keuleers et al. (2010, 2012) did not counterbalance the blocks across participants, the variability due to practice effects was not removed from the Dutch Lexicon Project and from the British Lexicon Project data, which

led the authors to further recommend the use of z-scores and/or the introduction of time-specific variables as covariates in the statistical analyses conducted on these chronometric data.<sup>1</sup>

### Predictors of visual word recognition and pronunciation of EP words

To explore the role played by the orthographic, phonological and semantic variables under study in visual recognition and pronunciation of EP words, multiple regression analyses were conducted considering the raw RTs and the accuracy rates of the items (words) that reached 67% of correct responses in both tasks ( $N = 1,911$ ) as dependent variables. Since raw RTs and zRTs produced similar reliability scores, we opted to use the raw RTs since the results from this dependent measure (expressed in ms) yielded a better interpretation of the effects. Notwithstanding, conducting the same analyses on the zRTs produced virtually the same results. Moreover, it is also worth noting that since several variables were highly correlated (see Table 3), which may increase multicollinearity problems in the regression analyses (see Cohen, Cohen, West, & Aiken, 2003), we first explored which of the word frequency, word length and word similarity measures was the best determinant of EP word performance. For that purpose, we conducted separate regression analyses and examined the amount of variance accounted for by each of them in both tasks (see Soares, Machado, et al., 2015; for a similar procedure). Only then, the role played by the semantic variables was examined, by conducting multiple regression analyses controlling for the effect of word frequency, word length, and word neighbour variables shown to be the best determinants of EP word processing in the previous analyses.

### Word frequency effects on visual word recognition and pronunciation of EP words

To analyze which of the word frequency measures accounted for higher percentages of variance, separate regression analyses were conducted for each of them. Table 4 presents the results ( $R^2$ ) of the analyses conducted both on the latency and accuracy data from the lexical decision and naming tasks. P-PAL<sub>pmwfr</sub>, SUBTLEX-PT<sub>pmwfr</sub> and SUBTLEX-PT<sub>CD</sub> counts were both log transformed ( $\text{Log}_{10}$ ) and squared ( $\text{Log}_{10}^2$ ) considering that Balota et al. (2004; see also Baayen et al., 2006) found that the relationship between the log-transformed frequency and word latencies is not completely linear. Log10 transformations were computed after summing 1 to the base value of each measure as suggested by Brysbaert and Diependaele (2013). P-PAL<sub>zipfr</sub>, SUBTLEX-

**Table 3.** Linear correlations between the full set of orthographic, phonological and semantic variables and the latency (RT in ms) and accuracy (%Acc) rates from the lexical decision (LDT) and naming (NAM) data.

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. LDT RT	–	–.591**	.475**	–.163**	–.263**	–.425**	–.288**	–.588**	–.455**	.365**	.365**	.319**	–.182**	.363**	.375**
2. LDT %Acc		–	–.160**	.052**	.118*	.282**	.069**	.270**	.167**	.141**	.099**	.072**	–.106**	.135**	.126*
3. NAM RT			–	–.096**	–.119**	–.160**	–.196**	–.357**	–.296**	.483**	.450**	.407**	–.289**	.478**	.465**
4. NAM %Acc				–	.057*	.076**	.076**	.128**	.120**	–.112**	–.128**	–.092**	.034ns	–.103**	–.120*
5. P-PAL <sub>pmwf</sub>					–	.737**	.511**	.477**	.559**	–.109*	–.113*	–.099*	.054*	–.106*	–.107*
6. P-PAL <sub>Zipf</sub>						–	.398**	.569**	.528**	–.063*	–.088*	–.060*	.014ns	–.066*	–.066*
7. SUBTLEX-PT <sub>pmwf</sub>							–	.629**	.826**	–.236**	–.239**	–.198**	.173**	–.230**	–.237**
8. SUBTLEX-PT <sub>Zipf</sub>								–	.852**	–.378**	–.364**	–.332**	.254**	–.364**	–.380**
9. SUBTLEX-PT <sub>CD</sub>									–	–.329**	–.319**	–.288**	.234**	–.319**	–.329**
10. $N_{\text{lett}}$										–	.893**	.797**	–.625**	.993**	.959**
11. $NO_{\text{syll}}$											–	.717**	–.556**	.889**	.898**
12. $OLD_{20}$												–	–	.787**	.778**
13. ON													–	–	–.603**
14. OUP														–	.949**
15. $N_{\text{phon}}$															–
16. $NP_{\text{syll}}$															
17. PN															
18. Front								–							
19. Central															
20. Back															
21. High															
22. Mid															
23. Low															
24. Rounded															–
25. Bilabial															
26. Labiodental															
27. Apico-dental															
28. Alveolar															
29. Palatal															
30. Velar															
31. Stop oral															
32. Stop nasal															
33. Fricative															
34. Lateral															
35. Voiced															
36. Paroxytone															
37. Oxytone															
38. Proparoxytone															
39. Imag															
40. Conc															
41. SubjFreq															
42. AoA															
43.Val															
44. Arou															
45.Dom															

(Continued)



Table 3. Continued.

Variables	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1. LDT RT	.360**	-.155**	.072**	.047**	-.030ns	.038ns	.063**	-.014ns	-.029ns	-.097**	-.054**	.090**	.004ns	-.047*	-.007ns
2. LDT %Acc	.108**	-.101**	.020ns	.010ns	.027ns	.043ns	-.014ns	.033ns	.027ns	.037ns	-.009ns	-.052**	-.043ns	.012ns	.007ns
3. NAM RT	.443**	-.269**	.136**	.059**	.073**	.061**	.133**	.051**	.073**	.142**	-.256**	-.122**	-.003ns	-.131**	.037ns
4. NAM %Acc	-.131**	.038ns	-.059**	-.067**	.015ns	-.064**	-.053*	.006ns	.015ns	.021ns	-.006ns	-.001ns	.009ns	.028ns	.056*
5. P-PAL <sub>pmwf</sub>	-.117**	.033ns	-.008ns	-.026ns	.006ns	-.011ns	-.032ns	.026ns	.006ns	.047*	.025ns	-.012ns	.019ns	-.002ns	-.040ns
6. P-PAL <sub>zipf</sub>	-.095**	-.004ns	.028ns	-.025ns	.018ns	.029ns	-.040ns	.041ns	.018ns	.049*	.025ns	-.041ns	-.011ns	.015ns	-.031ns
7. SUBTLEX-PT <sub>pmwf</sub>	-.232**	.150**	-.050*	-.041ns	.006ns	-.037ns	-.052*	.011ns	.006ns	.043ns	.053*	-.008ns	.022ns	.006ns	-.023ns
8. SUBTLEX-PT <sub>zipf</sub>	-.359**	.226**	-.053*	-.052*	.026ns	-.027ns	-.063**	.011ns	.026ns	.088**	.050*	-.054*	.018ns	.041ns	-.031ns
9. SUBTLEX-PT <sub>CD</sub>	-.312**	.207**	-.060**	-.050*	.020ns	-.032ns	-.065**	.010ns	.020ns	.076**	.057*	-.032ns	.028ns	.032ns	-.041ns
10. N <sub>lett</sub>	.892**	-.575**	.150**	.105**	-.027ns	.173**	.056*	.001ns	-.027ns	-.096**	-.104**	.048*	-.083**	-.076**	.027ns
11. MO <sub>syll</sub>	.959**	-.507**	.191**	.208**	.032ns	.211**	.173**	.051*	.032ns	-.144**	-.109**	.025ns	-.076**	-.098**	-.033ns
12. OLD <sub>20</sub>	.702**	-.696**	.173**	.088**	.035ns	.184**	.063**	.037ns	.035ns	-.083**	-.086**	.041ns	-.023ns	-.024ns	-.073**
13. ON	-.542**	.931**	-.149**	-.102**	-.068**	-.159**	-.090**	-.059**	-.068**	.084**	.090**	-.030ns	.020ns	-.013ns	.084**
14. OUP	.889**	-.569**	.147**	.101**	-.031ns	.169**	.052*	-.004ns	-.031ns	-.089**	-.101**	.049*	-.085**	-.078**	.025ns
15. N <sub>phon</sub>	.895**	-.551**	.089**	.096**	-.038ns	.140**	.020ns	.006ns	-.038ns	-.072**	-.093**	.071**	-.066**	-.094**	.022ns
16. NP <sub>syll</sub>	-	-.490**	.199**	.215**	.031ns	.223**	.177**	.045ns	.031ns	-.146**	-.116**	.034ns	-.078**	-.101**	-.043ns
17. PN	-	-	-.138**	-.112**	-.074**	-.149**	-.104**	-.060**	-.074**	.098**	.100**	-.045*	.042ns	-.021ns	.073**
18. Front	-	-	-	-.108**	-.049*	.604**	.111**	-.026ns	-.049*	-.142**	-.085**	-.132**	-.067**	-.043ns	-.141**
19. Central	-	-	-	-	-.070**	.253**	.725**	-.056*	-.070**	-.201**	-.120**	-.186**	-.094**	-.061**	-.199**
20. Back	-	-	-	-	-	.073**	.011ns	.787**	1.000**	-.092**	-.055*	-.085**	-.043ns	-.028ns	-.091**
21. High	-	-	-	-	-	-	-.119**	-.048*	.073**	-.170**	-.102**	-.158**	-.080**	-.052*	-.169**
22. Mid	-	-	-	-	-	-	-	-.052*	.011ns	-.186**	-.111**	-.172**	-.087**	-.057*	-.184**
23. Low	-	-	-	-	-	-	-	-	.787**	-.074**	-.045ns	-.069**	-.035ns	-.023ns	-.073**
24. Rounded	-	-	-	-	-	-	-	-	-	-.092**	-.055*	-.085**	-.043ns	-.028ns	-.091**
25. Bilabial	-	-	-	-	-	-	-	-	-	-	-.159**	-.245**	-.124**	-.081**	-.262**
26. Labiodental	-	-	-	-	-	-	-	-	-	-	-	-.147**	-.074**	-.048*	-.157**
27. Apico-dental	-	-	-	-	-	-	-	-	-	-	-	-	-.115**	-.075**	-.242**
28. Alveolar	-	-	-	-	-	-	-	-	-	-	-	-	-	-.038ns	-.123**
29. Palatal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-.080**
30. Velar	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
31. Stop oral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
32. Stop nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
33. Fricative	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
34. Lateral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
35. Voiced	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
36. Paroxytone	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
37. Oxytone	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
38. Proparoxytone	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
39. Imag	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
40. Conc	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
41. SubjFreq	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
42. AoA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
43.Val	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
44. Arou	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
45.Dom	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

(Continued)

Table 3. Continued.

Variables	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
1. LDT RT	.020ns	-.084**	-.051*	.012ns	-.008ns	-.014ns	.002ns	.021ns	-.163**	-.064*	-.554**	.558**	-.259**	.108*	-.152**
2. LDT %Acc	.002ns	.005ns	-.011ns	-.039ns	-.069*	.028ns	-.036ns	.019ns	.022ns	-.046*	.381**	-.127**	.133**	-.065ns	.122**
3. NAM RT	.432**	-.112**	-.490**	-.201**	-.180**	-.010ns	-.025ns	.065**	-.126**	-.062**	-.267**	.432**	-.154**	.142**	-.087ns
4. NAM %Acc	.062***	.008ns	.010ns	-.006ns	-.011ns	.007ns	-.006ns	-.001ns	.039ns	.030ns	.129**	-.129**	.067ns	-.017ns	.028ns
5. P-PAL <sub>pmwf</sub>	-.002ns	.050*	.016ns	-.031ns	.001ns	.057*	-.058*	.012ns	-.139**	-.142**	.351**	-.153**	.159**	-.083ns	.111*
6. P-PAL <sub>Zipf</sub>	-.011ns	.056*	.002ns	-.043ns	-.011ns	.115**	-.108**	.004ns	-.210**	-.236**	.432**	-.158**	.223**	-.062ns	.150**
7. SUBTLEX-PT <sub>pmwf</sub>	-.001ns	.071**	.039ns	-.028ns	.039ns	.019ns	.006ns	-.045*	.038ns	-.014ns	.392**	-.382**	.172**	-.070ns	.077ns
8. SUBTLEX-PT <sub>Zipf</sub>	-.019ns	.099**	.049*	-.035ns	.031ns	-.042ns	.063**	-.047*	.150**	.046*	.549**	-.621**	.194**	-.069ns	.092*
9. SUBTLEX-PT <sub>CD</sub>	-.023ns	.089**	.057*	-.022ns	.043ns	-.025ns	.048*	-.047*	.033ns	-.051*	.550**	-.555**	.199**	-.073ns	.108*
10. N <sub>lett</sub>	.049*	-.099**	-.112**	-.080**	-.154**	.042ns	-.060**	.041ns	-.309**	-.287**	-.120**	.576**	-.094*	.175**	.009ns
11. NO <sub>syll</sub>	-.065**	-.089**	-.125**	-.071**	-.154**	-.065**	-.001ns	.116**	-.278**	-.264**	-.108**	.529**	-.073ns	.153**	.025ns
12. OLD <sub>20</sub>	-.028ns	-.063**	-.063**	-.088**	-.121**	.132**	-.194**	.138**	-.251**	-.240**	-.118**	.523**	-.077ns	.157**	.006ns
13. ON	.047*	.055*	.068**	.071**	.100**	-.144**	.197**	-.123**	.226**	.230**	.078**	-.457**	.040ns	-.115*	-.020ns
14. OUP	.052*	-.094**	-.108**	-.085**	-.155**	-.011	-.016ns	.049*	-.290**	-.270**	-.113**	.557**	-.098*	.160**	.004ns
15. N <sub>phon</sub>	.053*	-.079**	-.097**	-.053*	-.117**	.064**	-.095**	.070**	-.296**	-.276**	-.120**	.574**	-.100*	.176**	.007ns
16. NP <sub>syll</sub>	-.062**	-.095**	-.128**	-.082**	-.165**	-.116**	.036ns	.135**	-.264**	-.248**	-.104**	.516**	-.085	.158**	.009ns
17. PN	.045*	.055*	.058*	.092**	.123**	-.151**	.198**	-.113**	.201**	.203**	.058*	-.418**	.027	-.082ns	-.011ns
18. Front	-.229**	-.081**	-.135**	-.086**	-.196**	.025ns	-.033ns	.019ns	-.123**	-.127**	-.008ns	.112**	-.079	.073ns	-.039ns
19. Central	-.323**	-.115**	-.190**	-.122**	-.277**	.027ns	-.023ns	-.004ns	-.043ns	-.052*	-.008ns	.060ns	.004ns	.062ns	-.010ns
20. Back	-.148**	-.052*	-.087**	-.056*	-.127**	.003ns	-.014ns	.021ns	.004ns	.011ns	.002ns	.004ns	.102*	-.031ns	.112*
21. High	-.274**	-.097**	-.161**	-.103**	-.235**	.042ns	-.063**	.046*	-.121**	-.130**	.012ns	.119**	-.027ns	-.029ns	.015ns
22. Mid	-.298**	-.106**	-.176**	-.112**	-.256**	.001ns	.013ns	-.026ns	-.025ns	-.031ns	-.025ns	.031ns	.014ns	.118**	-.001ns
23. Low	-.119**	-.042	-.070**	-.045*	-.102**	.024ns	-.029ns	.012ns	-.013ns	.003ns	.000ns	.029ns	.037ns	.028ns	.033ns
24. Rounded	-.148**	-.052*	-.087**	-.056*	-.127**	.003ns	-.014ns	.021ns	.004ns	.011ns	.002ns	.004ns	.102*	-.031ns	.112*
25. Bilabial	.321**	.395**	-.250**	-.160**	.140**	.009ns	-.020ns	.102**	.105**	.102**	.023ns	-.083*	.010ns	.004ns	-.044ns
26. Labiodental	-.255**	-.091**	.634**	-.096**	.006ns	-.020ns	.016ns	.004ns	.026ns	.004ns	.039ns	-.067ns	.015ns	.016ns	.078ns
27. Apico-dental	.153**	-.140**	.296**	-.148**	.017ns	-.060**	.035ns	.039ns	-.074**	-.082**	-.033ns	.085*	-.138**	.026ns	-.129**
28. Alveolar	-.199**	.243**	-.117**	.475**	.340**	.015ns	.002ns	-.030ns	.040ns	.023ns	.006ns	-.036ns	.033ns	-.059ns	.040ns
29. Palatal	-.130**	-.046*	.323**	-.049*	.099**	.027ns	-.012ns	-.024ns	.049*	.051*	.038ns	-.111**	.096*	-.083ns	.040ns
30. Velar	.349**	-.150**	-.248**	.302**	.021ns	.004ns	.024ns	-.053*	.022ns	.057*	-.026ns	-.021ns	.041ns	-.051ns	.041ns
31. Stop oral	-	-.243**	-.402**	-.258**	-.067**	-.042ns	.053*	-.026ns	.040ns	.067**	-.015ns	-.018ns	-.107*	.077ns	-.119**
32. Stop nasal	-	-	-.143**	-.092**	.415**	.017ns	-.034ns	.035ns	.046*	.043ns	.016ns	-.048ns	.001ns	-.043ns	-.060ns
33. Fricative	-	-	-	-.152**	-.097**	-.019ns	.009ns	-.017ns	.029ns	.009ns	.036ns	-.064ns	.074ns	-.087ns	.090*
34. Lateral	-	-	-	-	.439**	.027ns	-.005ns	-.037ns	.005ns	.001ns	-.024ns	-.022ns	.066ns	-.073ns	.100*
35. Voiced	-	-	-	-	-	-.008ns	.025ns	-.032ns	.033ns	.021ns	-.011ns	-.085*	.036ns	.002ns	.014ns
36. Paroxytone	-	-	-	-	-	-	-.854**	-.129**	-.119**	-.090**	-.015ns	.124**	.145**	-.022ns	.082ns
37. Oxytone	-	-	-	-	-	-	-	-.406**	-.406**	-.406**	.010ns	-.189**	-.125**	.009ns	-.072ns
38. Proparoxytone	-	-	-	-	-	-	-	-	-.051*	-.048*	.007ns	.139**	-.016ns	.022ns	-.004ns
39. Imag	-	-	-	-	-	-	-	-	-	.886**	.003ns	-.598**	.124**	-.188**	.009ns
40. Conc	-	-	-	-	-	-	-	-	-	-	-.057*	-.519**	.033ns	-.281**	-.024ns
41. SubjFreq	-	-	-	-	-	-	-	-	-	-	-	-.544**	.330**	-.210**	.298**
42. AoA	-	-	-	-	-	-	-	-	-	-	-	-	-.180*	.354**	-.045ns
43. Val	-	-	-	-	-	-	-	-	-	-	-	-	-	-.484**	.827**
44. Arou	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-.416**
45. Dom	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Notes: \*\* $p < .001$ , \* $p < .05$ ; ns: non-significant; P-PAL<sub>pmwf</sub>: per million word frequency, and P-PAL<sub>Zipf</sub>: Zipf scale word frequency as obtained from the P-PAL database (Soares, Iriarte, et al., 2018); SUBTLEX-PT<sub>pmwf</sub>: per million word frequency, SUBTLEX-PT<sub>Zipf</sub>: Zipf scale word frequency, and SUBTLEX-PT<sub>CD</sub>: Contextual Diversity word frequency as obtained from the SUBTLEX-PT database (Soares, Machado, et al., 2015); N<sub>lett</sub>: Number of letters, NO<sub>syll</sub>: Number of orthographic syllables, OLD<sub>20</sub>: Orthographic Levenshtein Distance, ON: Orthographic Neighbourhood size, OUP: Orthographic Uniqueness Point, N<sub>phon</sub>: Number of phonemes, NP<sub>syll</sub>: Number of phonological syllables, and PN: Phonological Neighbourhood size as obtained from the P-PAL database (Soares, Iriarte, et al., 2018). Phonological characteristics of the onsets (front, central, back, high, mid, low, rounding, bilabial, labiodental, apico-dental, alveolar, palatal, velar, stop oral, stop nasal, fricative, lateral and voiced) and word stress pattern (paroxytone, oxytone and proparoxytone) categorised dichotomously (0–1). Imag: Imageability; Conc: Concreteness; and SubjFreq: Subjective frequency obtained from the Minho Word Pool database (Soares et al., 2017); Val: Valence; Arou: Arousal; and Dom: Dominance obtained from Soares et al. (2012) norms; AoA: Age of Acquisition obtained from Cameirão and Vicente (2010) norms.

**Table 4.** Percentages of variance accounted by objective and subjective word frequency measures on the latency (RT in ms) and accuracy (%Acc) data from the lexical decision (LDT) and naming (NAM) tasks.

Word frequency measures	Behavioural measures	Task				
		LDT		NAM		
		RT	%Acc	RT	%Acc	
P-PAL	pmwf	$\text{Log}_{10}$	17.5	7.4	2.5	.06
		$\text{Log}_{10} + \text{Log}_{10}^2$	18.6	10.3	ns	ns
	Zipf	Zipf	18.1	8.0	2.6	.06
SUBTLEX-PT	pmwf	$\text{Zipf} + \text{Zipf}^2$	18.9	10.8	ns	ns
		$\text{Log}_{10}$	34.6	7.3	12.7	1.7
		$\text{Log}_{10} + \text{Log}_{10}^2$	35.6	9.0	ns	ns
	Zipf	Zipf	34.6	7.3	12.7	1.7
		$\text{Zipf} + \text{Zipf}^2$	35.6	9.0	ns	ns
CD	$\text{Log}_{10}$	37.2	8.8	13.8	1.8	
	$\text{Log}_{10} + \text{Log}_{10}^2$	37.4	10.1	ns	ns	
MWP	SubjFreq	SubjFreq	30.7	14.5	7.1	1.7
		$\text{SubjFreq} + \text{SubjFreq}^2$	31.7	22.2	ns	ns

Notes: P-PAL<sub>pmwf</sub>: per million word frequency, and P-PAL<sub>zipf</sub>: Zipf scale word frequency obtained from the P-PAL database (Soares, Iriarte, et al., 2018); SUBTLEX-PT<sub>pmwf</sub>: per million word frequency, SUBTLEX-PT<sub>zipf</sub>: Zipf scale word frequency, and SUBTLEX-PT<sub>CD</sub>: Contextual Diversity word frequency as obtained from the SUBTLEX-PT database (Soares, Machado, et al., 2015); SubjFreq: Subjective Frequency as obtained from the Minho Word Pool database (Soares et al., 2017).

PT<sub>Zipf</sub> and MWP<sub>SubjFreq</sub> were also squared to test for non-linear effects.

The inspection of Table 4 shows that regardless of the word frequency measure considered and of the dependent variable analyzed, word frequency accounted for higher percentages of variance in EP word recognition than in EP word pronunciation (in LDT it contributes, on average, for  $\cong 30\%$  of the variance in the RT data and for  $\cong 12\%$  of the variance in the accuracy data, while in NAM it only contributes, on average, for  $\cong 8.6\%$  of the variance in the RT data and for  $\cong 1.4\%$  of the variance in the accuracy data). In line with previous studies (e.g. Balota et al., 2004, 2007; Cortese & Khanna, 2007; Ferrand et al., 2011; Treiman et al., 1995; Yap et al., 2010; Yap & Balota, 2009), these results indicate that in EP, as in other languages, the (facilitative) effects of word frequency were larger in word/nonword discriminations than in speeded pronunciation. This suggests other variables might impact upon reading aloud more strongly. Secondly, and irrespective of the differences in the percentage of variance accounted for by each of the word frequency measures across tasks, subtitle word frequencies accounted for higher percentages of variance than written-texts word frequencies in the latency data from both tasks. Specifically, in LDT, SUBTLEX-PT counts explained  $\cong 17\%$  more of the variance in the RT data than P-PAL word counts, and in NAM, SUBTLEX-PT explained  $\cong 10\%$  more of the variance in the RT data than P-PAL counts. These results are consistent with previous findings of Soares, Machado, et al. (2015), and others

in lexical decision (e.g. Ferrand et al., 2010, 2017; Keuleers et al., 2010; Sze et al., 2014; Tse et al., 2017) and naming performance (e.g. Cai & Brysbaert, 2010; Cuetos et al., 2011). Moreover, they also demonstrate that the MWP<sub>SubjFreq</sub> measure was no better than SUBTLEX-PT word frequency measures in predicting the speed with which EP words were recognised and pronounced, explaining on average  $\cong 5\%$  less of variance in the latency data from LDT than the SUBTLEX-PT measures, and  $\cong 6\%$  less of variance in the latency data from NAM than the SUBTLEX-PT measures. Nevertheless, it is important to note that in the accuracy data from LDT, MWP<sub>SubjFreq</sub> accounted for  $\cong 13\%$  more variance than SUBTLEX-PT word frequency measures, though it still explained  $\cong 6\%$  less variance than the SUBTLEX-PT word frequency measures in the accuracy data from NAM (see Table 4). Yet, it is also worth noting that the MWP<sub>SubjFreq</sub> measure accounted for  $\cong 12\%$  more variance in RT and accuracy data from LDT, and for  $\cong 5\%$  more variance in the RT data from NAM than the P-PAL word frequency measures. The difference in the accuracy data from NAM is less expressive (1% more). Thus, the current findings do not support the superiority of the subjective frequency measure observed in previous studies (e.g. Balota et al., 2001, 2004; Cortese & Khanna, 2007) at least when the subjective frequency estimates were compared with the objective word counts drawn from subtitles. Note that in the studies conducted so far showing an advantage of subjective over objective word frequency measures in word processing, subjective frequency was compared with objective word counts drawn from written-texts and not from subtitles, shown to be a better determinant of reading performance in several languages including EP (e.g. Soares, Machado, et al., 2015; Sze et al., 2014; Tse et al., 2017; Yap et al., 2010). As Balota et al. (2001) pointed out, it is possible than when suboptimal word frequency measures are used (see Soares, Machado, et al., 2015 for a discussion of the limitations of word counts drawn from written-texts as books and periodicals), other variables, such as subjective frequency, might “fill in the gap”. However, when more reliable word measures are used (as subtitles counts), the predictive power of these variables (subjective frequency) could be largely reduced, as it seems to be the case in our data. These findings place the subjective word frequency measure in between the two other objective word frequency measures used, and strongly advise the use of subjective word frequency estimations when only objective word frequency measures drawn from written-text corpus are available. However, if subtitle word counts are available, they should be preferred over the subjective word estimates as they seem to



represent a good proxy of the relative exposure/use of words in a given language.

Furthermore, although SUBTLEX-PT measures accounted for higher percentages of variance than P-PAL measures in the speed with which EP words were recognised and pronounced, the difference in the percentage of variance accounted for by each of these word frequency measures was lower (and not higher) in naming than in lexical decision performance ( $\cong 17\%$  in LDT and  $\cong 10\%$  in NAM). These results suggest that the pronunciation of EP words was much less sensitive to the number of times a word occurs in a language, even when considering that subtitle word counts corresponded to the transcriptions of social interactions that people made in different situations. Anyway, because the SUBTLEX-PT<sub>CD</sub> measure was shown to be the best predictor of EP word processing times in both tasks, we decided to use this word frequency measure in the subsequent analyses to explore the relative contribution that the other orthographic, phonological and semantic variables play in visual recognition and pronunciation of EP words, once the role of word frequency was partialled out.

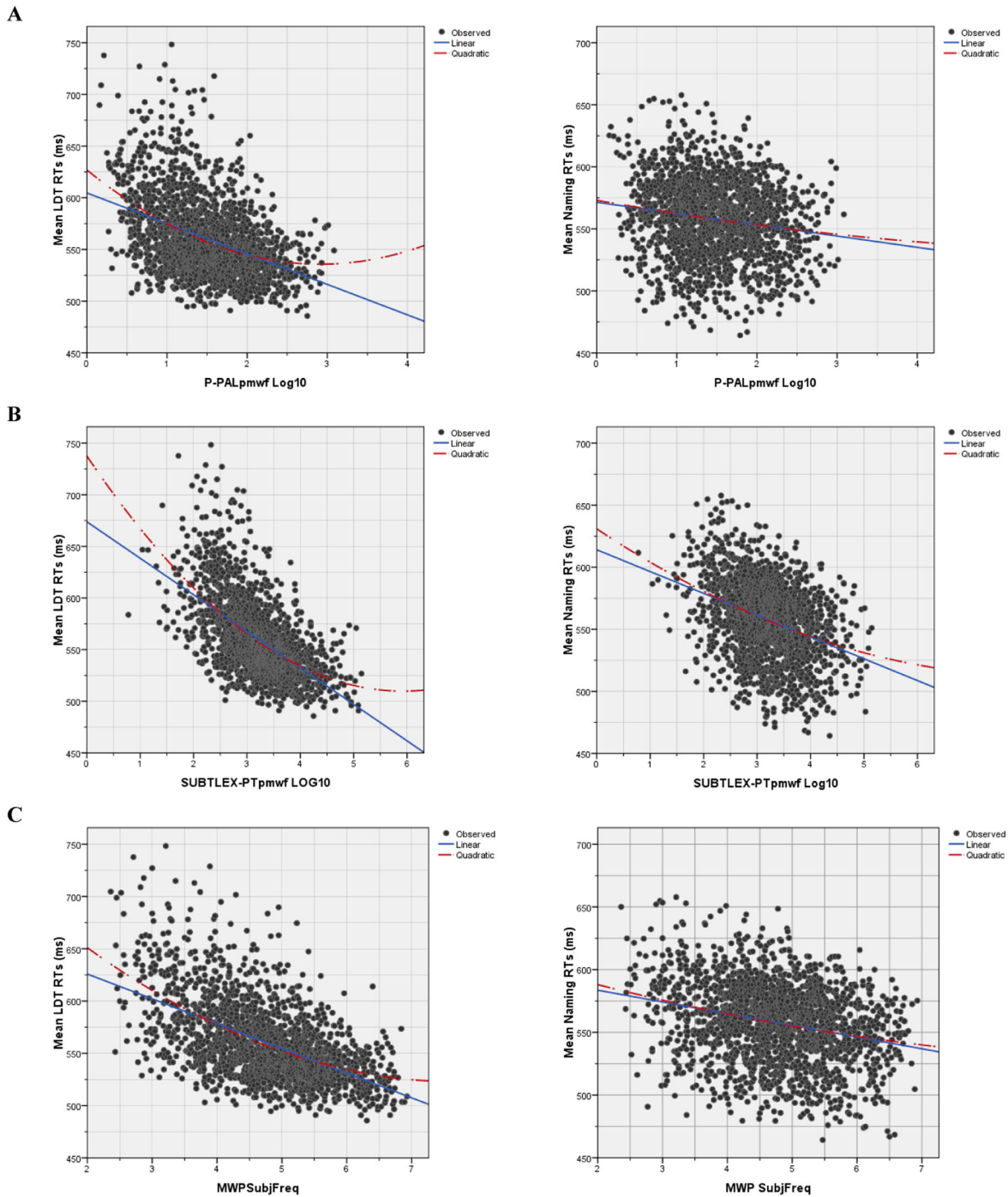
Another finding that should be highlighted from the analysis of Table 4 is the fact that the relationship between word frequency and latency and accuracy was better captured by a linear function in naming, but by a non-linear (quadratic) function in lexical decision. Indeed, in the LDT data, when the squared values of any of the word frequency measures were added to the regression equation as predictor, the percentage of variance explained increased significantly ( $F$  change, all  $ps < .001$ ) both in the RT (P-PAL<sub>pmwf</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 217.652$ ,  $p < .001$ , P-PAL<sub>zipf</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 222.987$ ,  $p < .001$ , SUBTLEX-PT<sub>pmwf</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 527.592$ ,  $p < .001$ , SUBTLEX-PT<sub>CD</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 526.231$ ,  $p < .001$ , SUBTLEX-PT<sub>zipf</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 570.520$ ,  $p < .001$ , MWP<sub>SubjFreq</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 442.941$ ,  $p < .001$ ), and accuracy data (P-PAL<sub>pmwf</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 109.030$ ,  $p < .001$ , P-PAL<sub>zipf</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 115.013$ ,  $p < .001$ , SUBTLEX-PT<sub>pmwf</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 94.906$ ,  $p < .001$ , SUBTLEX-PT<sub>CD</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 107.484$ ,  $p < .001$ , SUBTLEX-PT<sub>zipf</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 94.754$ ,  $p < .001$ , MWP<sub>SubjFreq</sub>  $\text{Log}_{10} + \text{Log}_{10}^2$ ,  $F(1, 1910) = 272.582$ ,  $p < .001$ ). This finding agrees with the results observed in other languages (e.g. Baayen et al., 2006; Balota et al., 2004; Brysbaert & New, 2009; Keuleers et al., 2010). However, in the naming data, the  $R^2$  for the linear and quadratic components was virtually the same. Hence, introducing the squared values of any of the word frequency measures as predictors in the analyses did not significantly change the percentage of variance

explained both in the latency (P-PAL<sub>pmwf</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 48.640$ ,  $p < .001$ , P-PAL<sub>zipf</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 50.049$ ,  $p < .001$ , SUBTLEX-PT<sub>pmwf</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 278.288$ ,  $p < .001$ , SUBTLEX-PT<sub>CD</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 305.397$ ,  $p < .001$ , SUBTLEX-PT<sub>zipf</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 278.315$ ,  $p < .001$ , MWP<sub>SubjFreq</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 145.963$ ,  $p < .001$ ), and accuracy data (P-PAL<sub>pmwf</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 11.040$ ,  $p < .01$ , P-PAL<sub>zipf</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 11.266$ ,  $p < .01$ , SUBTLEX-PT<sub>pmwf</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 33.516$ ,  $p < .001$ , SUBTLEX-PT<sub>CD</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 35.169$ ,  $p < .001$ , SUBTLEX-PT<sub>zipf</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 33.215$ ,  $p < .001$ , MWP<sub>SubjFreq</sub>  $\text{Log}_{10}$ ,  $F(1, 1910) = 32.877$ ,  $p < .001$ ). Figure 2 displays the scatter plots of the mean items latency (RT in ms) obtained from the lexical decision and naming tasks.

Because the SUBTLEX-PT<sub>CD</sub> accounted for much less variance in NAM than in LDT performance ( $\Delta R^2 = 23.6\%$ ), and also because previous studies (e.g. Balota et al., 2004; Chateau & Jared, 2003; Cortese & Khanna, 2007; Davies et al., 2013; Treiman et al., 1995; Yap & Balota, 2009) demonstrated that words' phonological characteristics accounted for significant amounts of variance in speeded pronunciation, we conducted additional regression analyses considering the phonological properties (onsets and stress-pattern) of the EP words used in both tasks as predictors. Table 5 presents the regression coefficients obtained from the lexical decision and naming data (RTs and %Accs).

As in other languages (e.g. Balota et al., 2004; Chateau & Jared, 2003; Cortese & Khanna, 2007; Davies et al., 2013; Spieler & Balota, 2000; Treiman et al., 1995; Yap & Balota, 2009), the phonological properties of the EP words were quite powerful in predicting the speed with which EP words were pronounced. Indeed, altogether they accounted for 43.2% of the variance in the RT data,  $F(16, 1910) = 90.135$ ,  $p < .001$ , although they only accounted for 1.5% of the variance in the accuracy data,  $F(16, 1910) = 1.749$ ,  $p = .033$ . Note that the SUBTLEX-PT<sub>CD</sub> word frequency measure, which was demonstrated to be the best determinant of EP word processing in the previous analyses, only accounted for 13.8% of the variance in the RT naming data ( $\Delta R^2 = 29.4\%$ ), thus leaving much room for other variables to exert their influence. However, in the LDT, coding the onsets and the stress pattern of the EP words only accounted for 3.5% of the variance in the RT data,  $F(16, 1910) = 4.305$ ,  $p < .001$  – the SUBTLEX-PT<sub>CD</sub> measure accounted for 37.4% of the variance in the RT LDT data ( $\Delta R^2 = 33.9\%$ ) (see Table 4).

Moreover, from Table 5 it is also possible to observe that all the phonological features significantly contributed to the speed with which EP words were pronounced, except the apico-dental and oxytone features. From the



**Figure 2.** A: Scatterplots depicting the linear and the quadratic functions of the  $P\text{-PAL}_{\text{pmwf}}$  (log transformed) measure obtained from Soares, Iriarte, et al. (2018) on the latency (RT in ms) data from the lexical decision (LDT) and naming (NAM) tasks. B: Scatter plots depicting the linear and the quadratic functions of the  $\text{SUBTLEX-PT}_{\text{pmwf}}$  (log transformed) measure obtained from Soares, Machado, et al. (2015) on the latency (RT in ms) data from the lexical decision (LDT) and naming (NAM) tasks. C: Scatter plots depicting the linear and the quadratic functions of the  $\text{MWP}_{\text{SubjFreq}}$  measure obtained from Soares et al. (2017) on the latency (RT in ms) data from the lexical decision (LDT) and naming (NAM) tasks.

predictors that reached statistical significance, all contributed positively to the speed with which EP words were pronounced, except voicing and velar. Hence, presenting each of these onset features (particularly stop oral, mid, high and stop nasal) contributed to a

significant increase of the time needed to pronounce EP words. This could be related either to the sensitivity of the voice-key to be more easily triggered by these phonetic features, and/or to the difficulty with which these phonological codes were implemented during

**Table 5.** Raw Regression Coefficients (Standardised Regression Coefficients in brackets) of the phonological characteristics of the EP words on the latency (RT in ms) and accuracy (%Acc) data from the lexical decision (LDT) and naming (NAM) tasks.

Predictors	Task	LDT		NAM	
		RT	%Acc	RT	%Acc
First-phoneme characteristics (onsets)	Front	7.503 (.050)ns	-.155 (-.010)ns	16.734 (.137)***	-.020 (-.005)ns
	High	15.612 (.121)**	.070 (.005)ns	34.000 (.322)***	-.278 (-.074)ns
	Mid	20.579 (.171)***	-.670 (-.052)ns	47.396 (.480)***	-.220 (-.062)ns
	Low	23.637 (.087)*	.609 (.021)ns	34.036 (.153)***	-.372 (-.047)ns
	Rounded	-13.438 (-.060)ns	-.250 (-.011)ns	15.513 (.085)**	.383 (.059)ns
	Labiodental	6.237 (.046)ns	-.434 (-.030)ns	14.937 (.134)***	-.013 (-.003)ns
	Apico-dental	15.434 (.156)***	-.716 (-.067) <sup>a</sup>	-.713 (-.009)ns	-.024 (-.008)ns
	Alveolar	8.162 (.048)ns	-.763 (-.042)ns	23.000 (.166)***	.124 (.025)ns
	Palatal	.600 (.002)ns	.201 (.007)ns	17.245 (.084)**	.248 (.034)ns
	Velar	4.750 (.050)ns	-.201 (-.020)ns	-5.759 (-.074)**	.098 (.035)ns
	Stop oral	8.443 (.108) <sup>a</sup>	-.071 (-.008)ns	59.717 (.931)***	.055 (.024)ns
	Stop nasal	-2.690 (-.019)ns	.377 (.025)ns	27.908 (.240)***	.103 (.025)ns
	Lateral	5.850 (.043)ns	.040 (.003)ns	19.956 (.180)***	-.028 (-.007)ns
	Voiced	4.156 (.051)ns	-.682 (-.078)*	-7.108 (-.106)***	-.145 (-.061) <sup>a</sup>
	Word stress pattern	Oxytone	-.562 (-.006)ns	.249 (.025)ns	.477 (.006)ns
Proparoxytone		3.739 (.023)ns	.317 (.018)ns	10.623 (.079)***	.015 (.003)ns
	$R^2$	.035	0.13	.432	.015

Notes: \*\*\* $p < .001$ ; \*\* $p < .01$ ; \* $p < .05$ ; <sup>a</sup> marginally significant; ns: nonsignificant; Phonological characteristics of the onsets (front, high, mid, low, rounding, labiodental, apico-dental, alveolar, palatal, velar, stop oral, stop nasal, lateral and voiced) and word stress pattern (oxytone and proparoxytone) categorised dichotomously (0–1). Regression coefficients for the central, back, bilabial, fricative, and proparoxytone features are not presented because they were not calculated due to collinearity issues.

articulation (see Kessler, Treiman, & Mullennix, 2002; for details). In the naming accuracy data, voicing was the only variable that contributed significantly and negatively to the accuracy with which EP words were pronounced, though only at a marginally significant level ( $p = .081$ ).

In latency data from LDT, only the tongue height (high, mid, and low) and the apico-dental and the stop oral features contributed significantly and positively to the speed with which EP words were recognised (though, in this last case, the coefficient was only marginally significant,  $p = .055$ ). In the LDT accuracy data, voicing and apico-dental were the only phonetic features that contributed significantly and negatively to the accuracy of word/nonword discriminations (though, in the last case, only at a marginally significant level,  $p = .069$ ). Nevertheless, and despite differences in the size of the individual regression coefficients observed across tasks, the direction of the effects was identical in lexical decision and naming performance. This suggests that the articulatory and/or phonological processes involved in EP word pronunciation also contribute (though to a lower extent) to EP visual word recognition. Hence, the effects of these variables were considered in the subsequent analyses.

### Word length effects on visual word recognition and pronunciation of EP words

To explore which of the word length measures (letters, phonemes, orthographic and phonological syllables) produced larger effects on visual word recognition and pronunciation of EP words, separate regression analyses

were conducted as in the previous analyses to avoid multicollinearity problems (Cohen et al., 2003), since word length measures were strongly correlated (see Table 3). The results ( $R^2$ ) of the regression analyses are presented in Table 6. The raw length values, along with its squared value, were entered as predictors in the regression equations of each analysis to explore nonlinear effects as in previous studies (e.g. Ferrand et al., 2010, 2017; New et al., 2006).

As shown in Table 6, all word length measures produced larger effects on naming than on lexical decision performance. Indeed, in NAM, they accounted for  $\cong 21.5\%$  of the variance in the RT data, and for  $\cong 1.8\%$  of

**Table 6.** Percentage of variance accounted for by word length measures in number of letters and orthographic and phonological syllables on the latency (RT in ms) and accuracy (%Acc) data from the lexical decision (LDT) and naming (NAM) tasks.

Word length measures	Behavioural measures	Task			
		LDT		NAM	
		RT	%	RT	%
Number of letters	$N_{\text{lett}}$	13.4	2.0	23.3	1.2
	$N_{\text{lett}} + N_{\text{lett}}^2$	14.7	2.6	23.8	1.5
Number of Phonemes	$N_{\text{phon}}$	14.0	1.6	21.6	1.4
	$N_{\text{phon}} + N_{\text{phon}}^2$	15.1	2.0	22.0	1.7
Number of syllables	Orthographic				
	$NO_{\text{syll}}$	13.3	1.0	20.2	1.6
	$NO_{\text{syll}} + NO_{\text{syll}}^2$	14.0	1.3	20.5	1.9
Phonological	$NP_{\text{syll}}$	13.0	1.2	19.6	1.7
	$NP_{\text{syll}} + NP_{\text{syll}}^2$	13.5	1.5	19.8	2.0
	$NP_{\text{syll}}^2$				

Notes:  $N_{\text{lett}}$ : Number of letters,  $N_{\text{phon}}$ : Number of phonemes,  $NO_{\text{syll}}$ : Number of orthographic syllables, and  $NP_{\text{syll}}$ : Number of phonological syllables as obtained from P-PAL database (Soares, Iriarte, et al., 2018).

the variance in the accuracy data, while in LDT they only accounted for  $\cong 14.3\%$  of the variance in the RT data ( $\Delta R^2 = 7.2\%$ ), though for  $\cong 1.9\%$  of the variance in the accuracy data ( $\Delta R^2 = -0.1\%$ ). These findings are in line with the results observed in previous large-scale studies conducted in deep (e.g. Balota et al., 2004; Cortese & Khanna, 2007; Yap & Balota, 2009) and shallow orthographies (e.g. Cuetos & Barbón, 2006; Davies et al., 2013; Wilson et al., 2013; Yap et al., 2010), but not with the findings of Lima and Castro (2010) in a previous factorial study with EP participants. Indeed, as mentioned before, Lima and Castro (2010) found reliable length effects on naming performance only when words and nonwords were presented in mixed lists, and on lexical decision only for words differing in two letters. These findings led the authors to conclude that in EP the use of the grapheme-phoneme strategy for phonological recodification is not as predominant as in other shallow orthographies. However, our results clearly indicate that when a higher and more diversified set of EP words is used, reliable word length effects are observed not only in reading aloud, as observed in languages with higher levels of orthographic-phonologic consistency (e.g. Barca et al., 2002; Burani et al., 2007; Cuetos & Barbón, 2006; Davies et al., 2013; González-Nosti et al., 2014; Wilson et al., 2013; Yap et al., 2010), but also in word/nonword discriminations. Note that the word length effects observed on the latency data from LDT were larger than the effects previously observed by, for example, Ferrand et al. (2010, 2011) or Keuleers et al. (2010) in the lexical decision time data from the French Lexicon Project and the Dutch Lexicon Project, respectively. As word length effects are considered an index of the engagement of the phonological route in word processing, these results also suggest that EP skilled readers rely strongly on the grapheme-phoneme conversion strategy when processing EP words, even in tasks that are primarily orthographic, as the LDT. Nevertheless, it is worth noting that word frequency effects were larger in lexical decision than in naming performance, hence indicating that in LDT the serial sub-lexical route of processing seems to operate simultaneously with higher-order lexical processes to allow a more efficient EP visual word recognition.

Moreover, and despite differences in the percentage of variance accounted for by each of the word length measures under analysis across tasks, it is also important to highlight that each of them produced comparable effects both in lexical decision and naming performance. Nonetheless, in the RT data from LDT the number of phonemes accounted for a slightly higher percentage of variance than both the number of letters ( $\Delta R^2 = 0.4\%$ ) and the number of orthographic ( $\Delta R^2 = 1.1\%$ )

and phonological ( $\Delta R^2 = 1.6\%$ ) syllables; and in the RT data from NAM the number of letters accounted for a slightly higher percentage of variance than both the number of phonemes ( $\Delta R^2 = 1.8\%$ ) and the number of orthographic ( $\Delta R^2 = 3.3\%$ ) and phonological ( $\Delta R^2 = 4\%$ ) syllables (see Table 6). In the accuracy data from LDT, however, the number of letters accounted for a slightly higher percentage of variance than both the number of phonemes ( $\Delta R^2 = 0.6\%$ ) and the number of orthographic ( $\Delta R^2 = 1.3\%$ ) and phonological ( $\Delta R^2 = 1.1\%$ ) syllables, and in the accuracy data from NAM the number of phonological syllables accounted for a slightly higher percentage of variance than both the number of orthographic syllables ( $\Delta R^2 = 0.1\%$ ) and the number of letters ( $\Delta R^2 = 0.5\%$ ) and phonemes ( $\Delta R^2 = 0.3\%$ ). Thus, contrary to what was previously observed in the English language (e.g. New et al., 2006; Yap & Balota, 2009), in both tasks single size units (letters/phonemes) accounted for a slightly higher percentage of variance in RT and accuracy data than syllable size units, even though syllable units (phonological/orthographic) accounted for a slightly higher percentage of variance than single size units (letters/phonemes) in the accuracy data from NAM. However, the similarities in the results observed across word length measures in both tasks suggest that both smaller (letters/phonemes) and larger (syllables) size units are activated during EP word processing. This provides further support for the notion that the orthographic and phonological codes are highly interconnected in intermediate-depth languages, and that both letters/phonemes and syllable size units play a functional role in visual word recognition and pronunciation, as demonstrated in previous EP studies using factorial designs (e.g. Campos, Oliveira, et al., 2018; Campos, Soares, et al., 2018; Lima & Castro, 2010; Pureza et al., 2016). These findings also support the claims of the grain size theory (Goswami & Ziegler, 2006; Ziegler & Goswami, 2005), namely that in nonshallow orthographies as EP, word processing involves the use of multiple phonological recoding units.

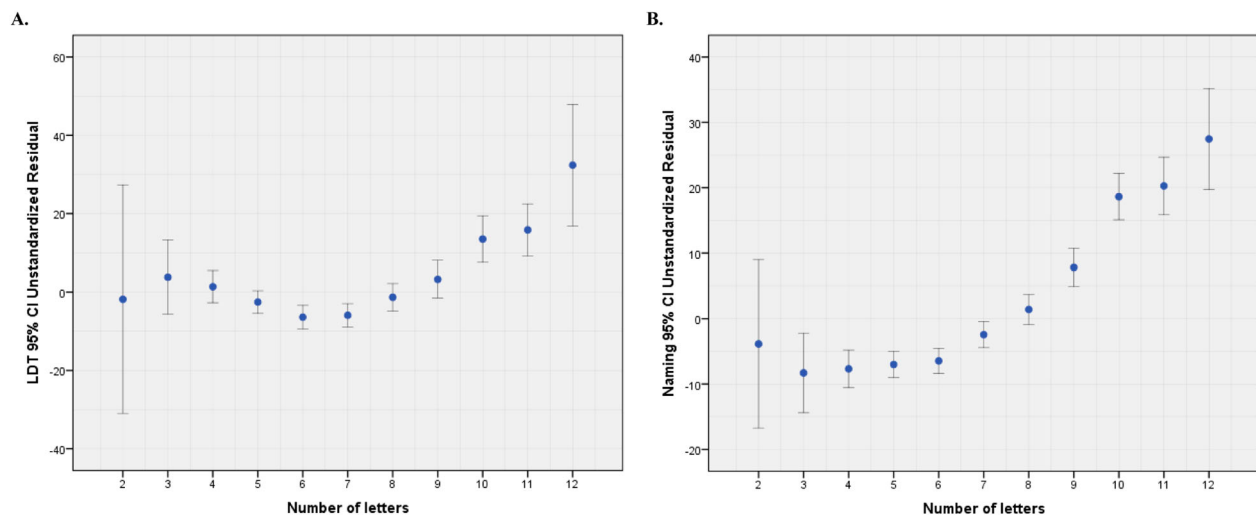
Another relevant finding is the fact that in both tasks word length effects were better captured by a nonlinear (quadratic) than by a linear function (see Table 6). Still, it is worth noting that differences in the percentage of variance accounted for by each of the word length measures considered when the squared valued were added to the regression equations were larger in the LDT ( $F$  change, all  $ps < .001$ ), both in the latency ( $N_{\text{lett}} + N_{\text{lett}^2}^2$   $F(1, 1910) = 164.169$ ,  $p < .001$ ,  $N_{\text{phon}} + N_{\text{phon}^2}^2$   $F(1, 1910) = 170.134$ ,  $p < .001$ ,  $NO_{\text{syll}} + NO_{\text{syll}^2}^2$   $F(1, 1910) = 155.291$ ,  $p < .001$ ,  $NP_{\text{syll}} + NP_{\text{syll}^2}^2$   $F(1, 1910) = 148.93$ ,  $p < .001$ ) and accuracy data ( $N_{\text{lett}} + N_{\text{lett}^2}^2$   $F(1, 1910) = 25.026$ ,  $p < .001$ ,  $N_{\text{phon}} + N_{\text{phon}^2}^2$   $F(1, 1910) = 19.379$ ,

$p < .001$ ,  $NO_{Syll} + NO_{Syll}^2$ ,  $F(1, 1910) = 12.954$ ,  $p < .001$ ,  $NP_{Syll} + NP_{Syll}^2$ ,  $F(1, 1910) = 14.948$ ,  $p < .001$ ), than in the NAM performance ( $F$  change, all  $ps < .05$ ), both in the latency ( $N_{lett} + N_{lett}^2$ ,  $F(1, 1910) = 298.281$ ,  $p < .001$ ,  $N_{phon} + N_{phon}^2$ ,  $F(1, 1910) = 268.746$ ,  $p < .001$ ,  $NO_{Syll} + NO_{Syll}^2$ ,  $F(1, 1910) = 245.649$ ,  $p < .001$ ,  $NP_{Syll} + NP_{Syll}^2$ ,  $F(1, 1910) = 235.851$ ,  $p < .001$ ) and accuracy data ( $N_{lett} + N_{lett}^2$ ,  $F(1, 1910) = 14.531$ ,  $p < .001$ ,  $N_{phon} + N_{phon}^2$ ,  $F(1, 1910) = 16.889$ ,  $p < .001$ ,  $NO_{Syll} + NO_{Syll}^2$ ,  $F(1, 1910) = 18.949$ ,  $p < .001$ ,  $NP_{Syll} + NP_{Syll}^2$ ,  $F(1, 1910) = 19.669$ ,  $p < .001$ ). These results suggest that the word length effects observed in EP are not monotonically defined, and that the nature and size of the effect change as a function of the number of letters, phonemes or syllables in the stimulus, as previously observed by New et al. (2006) and Yap and Balota (2009) for the English language, and by Ferrand et al. (2010, 2017) for the French language. Notwithstanding, contrary to New et al. (2006), robust nonlinear effects were observed both for single size (letters), and larger (syllables) size units, which suggests that, in EP, the nature of the relationship between word length and the speed/accuracy with which EP words are recognised/pronounced is not modulated by the type of word length measure considered.

To further explore the nature of this nonlinear relationship, we saved the residuals of the regression analyses considering the phonological properties of the EP words (onset and stress pattern) and the SUBTLEX-PT<sub>CD</sub> measure as predictors, and the RTs from each task as dependent variable, following the same procedure as Ferrand et al. (2010, 2017). Then, we analyzed the distribution of the mean value of the residuals as a function of the number of letters.<sup>2</sup> Figure 3 illustrates the results.

As shown in Figure 3, the relationship between word latencies and word length is defined by a U-shaped function, particularly in the lexical decision data (panel A), as previously reported for the English (New et al., 2006) and French languages (Ferrand et al., 2010, 2017). Indeed, RTs decreased for short words (i.e. for words from three to six letters), remained stable for medium long words (i.e. for words from six to seven letters), and increased for long words (i.e. for words from eight to 12 letters). In naming performance (Figure 3, panel B), the U-shaped function is less pronounced and the relationship between word length and RTs resembles the J-shaped function found by Ferrand et al. (2011) in the LDT data from the Chronolex database. Pronunciation times decreased slightly for words from two to three letters (short words), stayed relatively stable for four to six letters (medium words), and increased sharply for words with more than seven letters (long words). Hence, facilitative word length effects were not generally observed in naming, and inhibitory word length effects occurred earlier and more steeply on word pronunciation than on visual word recognition. The nonlinear nature of the EP word length functions observed in our data might even justify why Lima and Castro (2010) found null effects in lexical decision data for EP words with four to five letters, and also the fact that reliable length effects in naming were only observed when EP words and nonwords were presented in blocked conditions (see Lima & Castro, 2010; for details).

Moreover, to explore whether the word length effects observed in the EP language is an artifact of the number of similar words in the lexicon (i.e. of the neighbourhood measures as the  $N$  metric and the OLD<sub>20</sub> measure), as



**Figure 3.** Effect of word length (in number of letters) when the effect of the phonological characteristics of the first phoneme, word stress pattern, and objective word frequency measure (SUBTLEX-PT<sub>CD</sub>) as obtained from Soares, Machado, et al. (2015) has been partialled out. Error bars indicate the Standard Errors (SEs) of the mean.

observed in French (Ferrand et al., 2010; see Introduction), three-step hierarchical regression analyses were conducted on the latency and accuracy data from both tasks. The phonological characteristics of the words (onsets and stress pattern), and SUBTLEX-PT<sub>CD</sub> measure were entered as predictors in Step 1. Then, the OLD<sub>20</sub> measure was entered as predictor in Step 2, as this neighbourhood measure was shown to be the most appropriate orthographic word similarity determinant of EP word processing. Indeed, the results of the separate regression analyses conducted on the latency and accuracy data from both tasks showed that OLD<sub>20</sub> accounted for 16.6% of the variance in the RT data from the NAM,  $F(1, 1910) = 378.827, p < .001$ , and for 10.2% of the variance in the RT data from LDT,  $F(1, 1910) = 216.993, p < .001$ . In the accuracy data, OLD<sub>20</sub> accounted for 0.8% of the variance in NAM,  $F(1, 1910) = 16.240, p < .001$ , and for 0.5% of the variance in LDT,  $F(1, 1910) = 9.992, p < .01$ . Conversely, the orthographic *N* metric (ON) only accounted for 8.4% of the variance in the RT data from NAM,  $F(1, 1910) = 174.521, p < .001$  ( $\Delta R^2 = 8.2\%$ ), and for 3.3% of the variance in the RT data from LDT,  $F(1, 1910) = 65.375, p < .001$  ( $\Delta R^2 = 6.9\%$ ). In the accuracy data, ON only accounted for 0.1% of the variance in NAM,  $F(1, 1910) = 2.183, p = .110$  ( $\Delta R^2 = 0.7\%$ ), and for 1.1% of the variance in LDT,  $F(1, 1910) = 21.816, p < .001$  ( $\Delta R^2 = -0.6\%$ ), though in the former case the model was only marginally significant. Similar results were obtained for the phonological *N* metric (PN, Luce & Pisoni, 1998), the analogue of the Coltheart et al. (1977) neighbourhood measure in the phonological domain, which accounted for 7.2% of the variance in the RT data from NAM,  $F(1, 1910) = 149.197, p < .001$  ( $\Delta R^2 = 9.4\%$ ), and for 2.4% of the variance in the RT data from LDT,  $F(1, 1910) = 47.082, p < .001$  ( $\Delta R^2 = 7.8\%$ ). In the accuracy data, PN accounted for 0.1% of the variance in NAM,  $F(1, 1910) = 2.719, p = .099$  ( $\Delta R^2 = 0.7\%$ ), and for 1% of the variance in LDT,  $F(1, 1910) = 19.703, p < .001$  ( $\Delta R^2 = -0.5\%$ ), though, once again, in the former case the model was only marginally significant. Hence, these findings indicate that in EP, similarly to English, the OLD<sub>20</sub> measure represents a better proxy estimation of word similarity than the classic *N* metrics of Coltheart et al. (1977) and Luce and Pisoni (1998), not only in lexical decision (as previously observed by Yap & Balota, 2009; Yarkoni et al., 2008) but also in naming performance. For this reason, this word similarity measure was added to the regression analyses. In Step 3, the squared value of the word length (in number of letters) variable was entered as predictor to examine the unique variance accounted for by this measure when all other variables previously shown to affect EP word processing (i.e. word frequency, characteristics of the

first phoneme, stress pattern, similarity with other words in the lexicon) were controlled for, as in New et al. (2006) and Ferrand et al. (2010) studies. Even though OLD<sub>20</sub> and number of letters are strongly correlated in the current ( $r = .797, p < .001$ , see Table 3) as in previous studies (for example, in the study of Ferrand et al., 2010; the correlation between the OLD<sub>20</sub> and the number of letters was .771), the inspection of the collinearity statistics obtained when both predictors were included in the analyses was satisfactory according to the limits defined by Hair, Anderson, Tatham, and Black (1995). Indeed, the Variance Inflation Factor (VIF) was 2.935 for the OLD<sub>20</sub> measure and 3.024 for the  $N_{\text{lett}}^2$  measure, and the tolerance index was .341 for the OLD<sub>20</sub> measure and 3.31 for the  $N_{\text{lett}}^2$  measure. Moreover, the Durbin-Watson values for the independence of errors was also satisfactory (near 2 in both measures).

The results of the three-step hierarchical regression analyses showed that, despite a decrease in the percentage of variance explained by word length in both tasks when OLD<sub>20</sub> entered as predictor, sizeable proportions of variance were still observed. Specifically, in the latency data from LDT, the word length effect decreased from 4.2% of unique variance when the phonological characteristics of the words and the SUBTLEX-PT<sub>CD</sub> measure were entered as predictors,  $F(19, 1910) = 74.045, p < .001$ , to 2.8% of unique variance when OLD<sub>20</sub> also was also entered as predictor,  $F(20, 1910) = 70.419, p < .001$  (see Table 7). In the latency data from NAM, a similar pattern of results was observed, even though the reduction was sharp: from 10.3% when the phonological features of the EP words and the SUBTLEX-PT<sub>CD</sub> measure were added as predictors,  $F(19, 1910) = 186.348, p < .001$ , to 4% when OLD<sub>20</sub> was also considered,  $F(20, 1910) = 177.370, p < .001$  (see Table 7). In the accuracy data from LDT, the percentage of variance accounted for by word length decreased from 7% of unique variance when the characteristics of the first phoneme, stress pattern and SUBTLEX-PT<sub>CD</sub> were entered as predictors,  $F(19, 1910) = 219.595, p < .001$ , to 5.2% when OLD<sub>20</sub> was additionally introduced in the analyses,  $F(20, 1910) = 20.924, p < .001$ . A similar pattern was also observed in the accuracy data from NAM, although in this case the reduction was less pronounced: from 0.7% of unique variance when the characteristics of the first phoneme, stress pattern and SUBTLEX-PT<sub>CD</sub> were entered as predictors,  $F(19, 1910) = 3.862, p < .001$ , to 0.5% of unique variance when the OLD<sub>20</sub> was added to the regression equation,  $F(20, 1910) = 3.671, p < .001$ . Hence, sizeable word length effects were still observed in the EP language when the variables shown to affect EP word processing in the previous analyses (first-phoneme, word stress, subtitle word frequency and

**Table 7.** Raw Regression Coefficients (Standardised Regression Coefficients in brackets) accounted for by the semantic variables on the latency (RT in ms) and accuracy (%Acc) data from the lexical decision (LDT) and naming (NAM) tasks.

Task Predictors	LDT		NAM	
	RT	%Acc	RT	%Acc
Step 1: Onset + Stress + SUBTLEX-PT <sub>CD</sub> . $R^2$	.385***	.101***	.549***	.030***
Step 2: OLD <sub>20</sub> . $R^2$	.399***	.129***	.612***	.032**
Step 3: $N_{\text{lett.}}^2$ . $R^2$	.413***	.158***	.651***	.036**
Step 4: Semantic variables. $R^2$	.494***	.217**	.666***	.041***
Imag. $\beta$ unstandardised (standardised)	−7.850 (−.212)***	.773 (.197)***	−3.139 (−.139)***	.004 (.004)ns
Conc. $\beta$ unstandardised (standardised)	5.051 (1.670)***	−.439 (−.135)**	2.585 (.104)**	−.004 (−.004)ns
SubjFreq. $\beta$ unstandardised (standardised)	−14.492 (−.339)***	1.291 (.282)***	−4.951 (−.141)***	.120 (.096)**
Step 1: Onset + Stress + SUBTLEX-PT <sub>CD</sub> . $R^2$	.415***	.062***	.561***	.036*
Step 2: OLD <sub>20</sub> . $R^2$	.441***	.073***	.618***	.036*
Step 3: $N_{\text{lett.}}^2$ . $R^2$	.487***	.108***	.672***	.041*
Step 4: Semantic variables. $R^2$	.513***	.119***	.677***	.042*
AoA. $\beta$ unstandardised (standardised)	5.477 (.228)***	−.291 (−.151)***	2.394 (.119)***	−.033 (−.046)ns
Step 1: Onset + Stress + SUBTLEX-PT <sub>CD</sub> . $R^2$	.478***	.125***	.564***	.082**
Step 2: OLD <sub>20</sub> . $R^2$	.498***	.137***	.605***	.082**
Step 3: $N_{\text{lett.}}^2$ . $R^2$	.504***	.166***	.644***	.087**
Step 4: Affective variables. $R^2$	.519***	.172**	.647***	.093**
Val. $\beta$ unstandardised (standardised)	−3.561 (−.167)**	−.013 (−.006)ns	−1.539 (−.090) <sup>a</sup>	.030 (.052)ns
Arou. $\beta$ unstandardised (standardised)	−1.092 (−.032)ns	−.168 (−.050)ns	−.734 (−.027)ns	.010 (.011)ns
Dom. $\beta$ unstandardised (standardised)	1.442 (.033)ns	.230 (.053)ns	1.697 (.049)ns	−.044 (−.038)ns

Notes: \*\*\* $p < .001$ ; \*\* $p < .01$ ; \* $p < .05$ ; <sup>a</sup> marginally significant; ns: nonsignificant; Onset: Phonological characteristics of the onsets; Stress: word stress pattern; SUBTLEX-PT<sub>CD</sub>: Contextual Diversity word frequency from the SUBTLEX-PT database (Soares, Machado, et al., 2015); OLD<sub>20</sub>: Orthographic Levenshtein Distance, and  $N_{\text{lett.}}$ : Number of letters obtained from the P-PAL database (Soares, Iriarte, et al., 2018); Imag: Imageability, Conc: Concreteness, and SubjFreq: Subjective Frequency obtained from the Minho Word Pool database (Soares et al., 2017); Val: Valence, Arou: Arousal, and Dom: Dominance obtained from Soares et al. (2012) norms; AoA: Age of Acquisition obtained from Cameirão and Vicente (2010) norms.

word similarity) were taken into account, even though the percentage of unique variance decreased significantly, as observed in other languages (e.g. Ferrand et al., 2010, 2017; New et al., 2006).<sup>3</sup> The sharper reduction observed in naming than in lexical decision performance is also consistent with previous studies showing that the speed/accuracy with which words are processed are more strongly affected by the number of similar words in the lexicon in speeded pronunciation than in word/nonword discriminations (e.g. Andrews, 1997; Balota et al., 2004; Cortese & Khanna, 2007; Ferrand et al., 2011, 2017; Yap & Balota, 2009). Taken together, these results provide further support for the view that EP skilled readers rely strongly on the serial sub-lexical route of processing when recognising and particularly when pronouncing EP words.

To test whether the word length effects observed in EP were due to a strong reliance on the serial sub-lexical route of processing, additional analyses were conducted in which the orthographic uniqueness point (OUP) measure used in recent megastudies (e.g. Ernestus & Cutler, 2015; Ferrand et al., 2017; Goh et al., 2016) was introduced in Step 2 of the three-step hierarchical regression mentioned above instead of the OLD<sub>20</sub> measure.<sup>4</sup> Indeed, if this was the case, word length effects should disappear, or at least should be strongly diminished (bear in mind that the OUP measures are taken as an index of the use of the serial left-to-right route of processing as mentioned in the Introduction). The results of these additional analyses supported this

assertion. Indeed, although the word length effect remained significant, its impact on EP word processing was greatly reduced particularly in naming performance. Specifically, in the RT data from NAM, results showed that the effect attributed to word length decreased from 4% when the OLD<sub>20</sub> measure was considered to only 1% when the OUP was used instead,  $F(20, 1910) = 177.005$ ,  $p < .001$ . In the RT data from LDT, the results also indicated that when the OUP measure was included, the word length effect decreased, though in a lower extent, i.e. from 2.8% to 2.2%,  $F(20, 1910) = 70.759$ ,  $p < .001$ . In the accuracy data from LDT, the percentage of variance accounted for by word length decreased from 5.2% to 1.8%,  $F(20, 1910) = 21.068$ ,  $p < .001$ , though in the accuracy data from NAM, the inclusion of the OUP measure led to an increase in the percentage of variance from .05% to 1%,  $F(20, 1910) = 4.209$ ,  $p < .001$ . Hence, the reduction in the percentage of variance accounted for by word length when the OUP measure was introduced as a predictor suggests that the word length effects observed in EP resulted effectively from a stronger reliance on the serial left-to-right sub-lexical route of processing, particularly in naming performance. Moreover, the fact that word length effects were less attenuated in lexical decision than in naming performance is also consistent with the view that serial and parallel strategies operate simultaneously to allow a more efficient EP word recognition, as mentioned before.

Furthermore, the results obtained with the OUP measure also revealed a positive (facilitation) impact of

the OUP measure both in latency and accuracy data from word/nonword discriminations (RT:  $\beta_{\text{unstandardised}} = 6.787$ ,  $\beta_{\text{standardised}} = .363$ ,  $p < .001$ ; ACC:  $\beta_{\text{unstandardised}} = 0.271$ ,  $\beta_{\text{standardised}} = .135$ ,  $p < .001$ ), and in the latency data from speeded pronunciation (RT:  $\beta_{\text{unstandardised}} = 7.314$ ,  $\beta_{\text{standardised}} = .478$ ,  $p < .001$ ). This indicates that EP words with an early OUP were recognised faster and more accurately than words with a late OUP, as observed in previous factorial studies conducted in other languages (e.g. Kwantes & Mewhort, 1999; Luce, 1986; Radeau et al., 1989; see however Izura et al., 2014; for inhibitory OUP effects in the English language), though not with findings recently reported by Ferrand et al. (2017) on visual and auditory lexical decision data obtained from the MEGALEX database. Nevertheless, it should be also noted that the facilitation OUP effect observed in the EP language also entailed a cost in naming performance, as EP words with an early OUP produced more errors than EP words with a late OUP ( $\beta_{\text{unstandardised}} = -0.056$ ,  $\beta_{\text{standardised}} = -.103$ ,  $p < .001$ ), which was probably due to a trade-off effect: as participants were faster responding to words with an early OUP, these words might also be more prone to mispronunciations. Note that the correlation coefficient between accuracy and RTs for words with early OUP in the dataset (i.e. for words with an OUP measure below 7 – the median value in the distribution) is  $r = -.598$ ,  $p < .001$ .

### **Semantic effects on visual word recognition and pronunciation of EP words**

As noted before, there is a considerable debate about the unique contribution of semantic to word processing in different languages, although studies analyzing the role that these variables play in EP word recognition and pronunciation are, to the best of our knowledge, inexistent. To explore this issue, several hierarchical regression analyses were conducted on latency and accuracy data from both tasks, considering the imageability, concreteness, subjective frequency, AoA, valence, arousal, and dominance ratings provided by the EP normative studies reported by Soares et al. (2017), Cameirão and Vicente (2010) and Soares et al. (2012) (see methodological section). Even though these ratings are available for imageability, concreteness and subjective frequency of the total pool of items, they are only available for a limited set of words in the case of AoA, valence, arousal and dominance affective dimensions (see Table 1). For that reason, three different sets of four-step hierarchical regressions analyses were conducted for the latency and accuracy data from both tasks. Furthermore, it is also worth noting that despite the strong correlation between imageability and concreteness in the current study ( $r = .886$ ,  $p < .001$ , see Table 3),

the inspection of the collinearity statistics in the latency and accuracy data showed acceptable values (imageability: VIF = 5.008, tolerance = .220; concreteness: VIF = 5.040, tolerance = .210; subjective frequency: VIF = 1.571, tolerance = .636). The Durbin-Watson statistics was also satisfactory (near 2). Consequently, both variables were kept in the first set of analyses reported in Table 7. In the same vein, we also opted to maintain the valence and the dominance affective measures in the third set of regression analyses presented in Table 7. Indeed, although they also present a high correlation ( $r = .827$ ,  $p < .001$ , see Table 3), the inspection of the collinearity statistics also showed acceptable values (valence: VIF = 5.008, tolerance = .220; arousal: VIF = 5.040, tolerance = .210; dominance: VIF = 1.571, tolerance = .636). The Durbin-Watson statistics was also satisfactory in both cases (near 2). Nonetheless, caution is advised when interpreting the coefficients of predictors like imageability and concreteness (or valence and dominance), because they are highly related ( $r_s > .80$ ).

In all hierarchical regressions analyses conducted to examine the contribution of these semantic variables in EP word processing, the phonological characteristics (onsets and stress pattern) and word frequency (SUBTLEX-PT<sub>CD</sub>) were entered as predictors in Step 1, followed by the OLD<sub>20</sub> measure in Step 2, and by the square of word length ( $N_{\text{lett}}^2$ ) in Step 3. Only in Step 4, the semantic variables were introduced in the analyses to examine the percentage of “extra” variance accounted for by each of them. The results obtained for both tasks are presented in Table 7.

As shown in Table 7, imageability, concreteness and subjective frequency produced significant semantic effects both on lexical decision and naming performance, though larger on word/nonword discriminations than on speed pronunciation, in accordance with previous studies conducted in other languages (e.g. Balota et al., 1991, 2001, 2004; Cortese & Schock, 2013; Cortese & Khanna, 2007; Cuetos & Barbón, 2006; Davies et al., 2013; Ferrand et al., 2011; Goh et al., 2016; González-Nosti et al., 2014; Kousta et al., 2011; Kuperman, 2015; Wilson et al., 2013). Specifically in lexical decision, imageability, concreteness and subjective frequency accounted for more 8.1% of unique variance in the RT data,  $F(22, 1910) = 83.645$ ,  $p < .001$ , and for more 5.9% of unique variance in accuracy data,  $F(22, 1910) = 23.716$ ,  $p < .001$ , when all the other surface, lexical and sub-lexical variables were partialled out. Altogether, these variables accounted for 49.4% of the variance in the LDT latency data, and for 21.7% of the variance in the LDT accuracy data. In naming performance, imageability, concreteness and subjective frequency only accounted for more 1.5% of unique variance in the



latency data,  $F(22, 1910) = 170.894$ ,  $p < .001$ , and only for more 0.5% of unique variance in the accuracy data,  $F(22, 1910) = 3.709$ ,  $p < .00$ , when the phonetic features of the onsets, stress pattern, word frequency, word similarity and word length were partialled out (see Table 7). Together, these variables accounted for 66.6% of the total variance in the naming latency data, and for 4.1% of the total variance in the naming accuracy data. Thus, as in other languages, the analysis of the similarities and differences between the regression models examined separately for each experimental task suggests that the semantic variables seem to impact more EP word/nonword discrimination than EP word pronunciation. The weaker semantic effects observed in the naming performance may be related to the fact that the retrieval of phonological information from an orthographic input in EP can rely strongly on the use of sub-lexical recoding strategies (i.e. grapheme-conversion rules) as mentioned before. Thereby, the semantic-to-orthographic/phonological feedback connections may impact EP word pronunciation in a lesser extent than EP word/nonword discriminations.

Furthermore, and regardless of differences in the percentage of variance accounted for by the semantic variables under analyses across tasks, it is worth noting that whereas imageability and subjective frequency contributed negatively to the speed with which EP words were recognised and pronounced, and positively to the accuracy with which EP words were recognised (in naming the effect did not reach statistical significance, see Table 7), concreteness contributed positively to the latency data in both tasks and negatively to the accuracy data in the lexical decision task (in the naming data the result was nonsignificant, see Table 7). Hence, on the one hand, the more imageable and the more familiar EP words were, the faster and the more accurate the responses they elicited, in line with previous studies conducted in other languages (e.g. Balota et al., 2001, 2004; Cortese & Khanna, 2007; Ferrand et al., 2011; Wilson et al., 2013; Yap & Balota, 2009); on the other hand, the more concrete EP words were, the slower and the less accurate responses they produced.

This inhibitory concreteness effect observed in the EP data is inconsistent with the vast amount of factorial studies conducted in other languages, showing that concrete words were recognised, named, and recalled more quickly and easily than abstract words (see Kousta et al., 2011; or Bonin et al., 2018; for recent reviews), and also with the recent results obtained by Goh et al. (2016) in their lexical decision and semantic categorisation mega-study conducted with spoken English words. Nevertheless, observing concreteness reverse effects is not entirely new in the literature. For example, Kousta et al.

(2011) found similar results in a study that aimed to disentangle the role of two dominant accounts (dual coding theory and context availability model) in explaining the advantage of concrete words over abstract words. As abstract words were more affectively valenced than concrete words, the authors proposed that the denser affective associations in abstract than concrete words may yield richer internal representations, thereby leading abstract words to be recognised faster and more accurately than concrete words, when all other objective (e.g. word frequency, word length, *ON*), and subjective variables (e.g. imageability, familiarity, *AoA*) known to affect word recognition are controlled for (see Kousta et al., 2011; for details; see also Bonin et al., 2018; for similar results in the French language). The negative relationship observed between concreteness and both objective and subjective word frequency measures was also proposed by Soares et al. (2017) as another potential explanation. Indeed, since abstract words tend to be linked to a wide range of contexts/situations than concrete words, this could explain why abstract words were rated with higher values of use in everyday life than concrete words in the Minho Word Pool data (see Soares et al., 2017; for a further discussion), and also the concreteness inhibitory effect observed in the current study. Indeed, although concreteness was not significantly correlated with valence, it correlated significantly and negatively both with subjective frequency and SUBTLEX-PT<sub>CD</sub> measures (see Table 3), thus providing further support for this argument. Notwithstanding, regardless of the direction of the results whose discussion is beyond the scope of this paper, they clearly demonstrate that imageability, concreteness and subjective frequency contribute with significant proportions of “extra” variance in the processing of EP words, particularly in LDT. Hence, these variables should not be disregarded when planning/conducting research with EP words, particularly those relying more strongly on the meaningfulness of the stimuli (e.g. LDT, semantic categorisation task).

Regarding *AoA*, the results of the second set of regression analyses conducted were quite surprising. Indeed, contrary to the vast amount of studies showing that *AoA* is one of the most powerful variables in predicting lexical decision and naming times (e.g. Cortese & Khanna, 2007; Cortese & Schock, 2013; Cuetos & Barbón, 2006; Davies et al., 2013; Ferrand et al., 2011; González-Nosti et al., 2014; Wilson et al., 2013), the current findings show that *AoA* only contributes with a modest percentage of unique variance both in lexical decision and particularly in naming performance. Specifically, in lexical decision, *AoA* accounted for more 2.6% of unique variance in the latency data,  $F(21, 816) = 39.800$ ,

$p < .001$ , and for more 1.1% of unique variance in accuracy data,  $F(21, 816) = 5.106$ ,  $p < .001$  after all other variables shown to affect EP word processing were partialled out (see Table 7). In naming, AoA contributed for more 0.7% of unique variance in the latency data,  $F(21, 816) = 79.915$ ,  $p < .001$ , and for more 0.1% of unique variance in the accuracy data,  $F(21, 816) = 1.670$ ,  $p < .05$  when the phonetic features of the onsets, stress pattern, word frequency, word similarity and word length measures were controlled for (see Table 7). Hence, although the earlier in life a EP word is acquired, the easier it is processed, as shown in other languages (e.g. Cortese & Khanna, 2007; Cortese & Schock, 2013; Cuetos & Barbón, 2006; Davies et al., 2013; González-Nosti et al., 2014; Wilson et al., 2013), the magnitude of the AoA effects observed in EP were smaller than in other languages.

The less relevant role that AoA seems to play in EP word processing may be related with two main factors. First, we used a word frequency measure (SUBTLEX-PT<sub>CD</sub>) that was shown to be a better determinant of reading performance than the word frequency measures obtained from written-text corpus and used in previous studies in which the role of AoA was tested (e.g. Cortese & Khanna, 2007; Cortese & Schock, 2013; Cuetos & Barbón, 2006; Davies et al., 2013; González-Nosti et al., 2014; Wilson et al., 2013). Second, the OLD<sub>20</sub> measure was used instead of the classic *N* metric adopted in the abovementioned studies, which was shown to be a better proxy estimation of word similarity not only in English (e.g. Yap & Balota, 2009; Yarkoni et al., 2008), but also in EP. Yet, although less expressively, the percentage of variance accounted for by AoA in the EP data is not negligible (note that in some of the studies mentioned above semantic variables did not account for more than 1% of the variance even when using suboptimal word frequency and word neighbourhood measures). Thus, AoA should be not neglected when planning research with EP verbal stimuli.

Finally, the results obtained for the third set of regression analyses showed that the affective content of EP words contributed with even lower percentages of variance than AoA ratings both in the RT and accuracy data from lexical decision and naming performance (see Table 7). Specifically, in LDT, the affective variables only accounted for more 1.5% of unique variance in the latency data,  $F(22, 480) = 22.432$ ,  $p < .001$ , and for more 0.6% of unique variance in accuracy data,  $F(22, 480) = 4.340$ ,  $p < .001$ , after all the other surface, lexical and sub-lexical variables shown to affect EP word processing have been partialled out (see Table 7). Altogether, these variables accounted for 51.9% of the variance in the LDT latency data, and for 17.2% of the variance in the LDT

accuracy data. In naming, the contribution of the affective variables was even less expressive: more 0.3% of unique variance in the latency data,  $F(22, 480) = 38.086$ ,  $p < .001$ , and more 0.6% of unique variance in the accuracy data,  $F(22, 480) = 1.978$ ,  $p < .01$  when first-phoneme characteristics, word stress, SUBTLEX-PT<sub>CD</sub> and OLD<sub>20</sub> were controlled for. Together, all the variables accounted for 64.7% of the total variance in the latency naming data, and for 9.3% of the variance in the accuracy naming data. Of note, from all the affective measures considered, only valence contributed significantly and negatively to the speed with which EP words were recognised – note that none of the affective variables produced a significant effect in naming performance (see Table 7). This shows that EP skilled readers were faster at recognising positively valenced words than negatively valenced words, which is consistent with the positivity bias observed in previous studies in EP (e.g. Pinheiro et al., 2017; Soares et al., 2012, 2013; Soares, Pinheiro, et al., 2015; Vasconcelos, Dias, Soares, & Pinheiro, 2017) as well as in other languages (e.g. Goh et al., 2016; Kuperman, 2015; Kuperman et al., 2012), even though not controlling for the full set of variables used in the present analyses.

### Summary of findings

The current findings showed that orthographic, phonological and semantic variables have a strong impact on EP word processing, though the magnitude and the nature of the effects were modulated by task demands. Overall, word frequency (SUBTLEX-PT<sub>CD</sub>) and word similarity (OLD<sub>20</sub>) seem to have a stronger impact on EP word/nonword discrimination than EP word pronunciation; phonological properties of the onsets, stress pattern and word length (both in the number of letters/phonemes and orthographic/phonological syllables) seem have a stronger impact on EP word pronunciation than EP word/nonword decisions; and semantic variables seem to have stronger impact on EP lexical decision than EP word pronunciation.

These results are in line with previous findings observed in deep (e.g. English, French) and in shallow orthographies (e.g. Spanish, Dutch), hence putting EP in-between the effects observed in those languages. For instance, the strong reliance on the sub-lexical route of processing, indexed by a robust nonlinear word length effect in all word length measures considered (even when all other variables shown to affect word processing were controlled for), brings EP closer to shallow- than to deep-orthographies. The current study also showed that EP skilled readers make use of multiple recoding size units in reading, which brings EP closer to deep- than to shallow-orthographies. Moreover,

as observed in other deep and shallow languages, word frequency represents a powerful predictor of the speed with which EP words were recognised and pronounced, particularly when CD measure drawn from subtitles were used. Critically, the advantage of this measure over all other word frequency measures tested was still observed when the recent Zipf word frequency was used, and also when the subjective frequency measure was taken into account.

Even though the results from the current study demonstrate that the CD measure drawn from subtitles is undeniably the best index of word frequency, the phonetic features of the first-phoneme clearly go beyond word frequency measures by accounting for the largest portion of variance in the speed of EP word pronunciation. These findings are consistent with those observed in other deep- and shallow-orthographies, and suggest that the phonetic features of the first-phoneme should not be neglected particularly in studies using tasks that rely more strongly on words' phonological information. Additionally, as in (American) English, but in contrast with French or Malay languages, the OLD<sub>20</sub> represented the best proxy estimate of EP word similarity since it accounted for the largest amounts of variance in EP word processing, particularly in naming performance. However, contrary to the results of other large-scale studies conducted in deep-orthographies, facilitative OUP effects were observed both in EP visual word recognition and particularly in EP word pronunciation. This result provides further evidence for the view that EP skilled readers rely strongly on the sub-lexical recoding strategy when processing EP words, though EP skilled readers also rely on the lexical route of processing in visual word recognition.

Finally, as in other deep and shallow orthographies, the current results showed that imageability, concreteness and subjective frequency contribute with a sizeable percentage of variance in EP word processing, particularly in lexical decision. However, if EP words with higher imageability and subjective frequency ratings produced faster recognition times, a reverse concreteness effect was observed, as recently observed in the English and French languages. AoA and affective variables contributed with a very low percentage of variance in EP word processing, particularly in speeded pronunciation. Among the affective variables, only valence reached statistical significance in lexical decision, confirming in EP the positivity bias observed in other languages.

## Conclusion

The current study examined the role that orthographic, phonological and semantic variables play in visual

recognition and pronunciation of EP words, by using the megastudy approach. Although a growing body of evidence has been obtained from large-scale studies conducted in deep and shallow orthographies, little is known about how these word properties affect word processing in intermediate-depth language such as EP. We found that the pattern of findings in EP is in-between the effects observed in deep and shallow languages. A theoretical implication of the current megastudy is that models of visual word recognition and word production should address these differences by including different parameters/weights that better fit the singularities of each language and/or orthography (e.g. see Perea, Winkler, & Gomez, 2018; for discussion). Researchers interested in conducting new analyses probing the role that different variables play in EP word processing, or comparing word processing in EP and in other deep and/or shallow languages, may freely download the behavioural data described in the current study as a supplemental archive or at <http://p-pal.di.uminho.pt/about/databases>.

## Notes

1. Even though we have not used repeated blocks aiming to directly test practice effects in participants' performance as Keuleers et al. (2010), the use of lexically similar experimental blocks (see Materials section) allowed us to examine practice effects across blocks even considering that they relied on a different subset of stimuli.
2. Even though the number of phonemes accounted for 0.4% more variance in the latency data from TDL latencies, we opted to use the number of letters in the analyses since this word length measure has been used in most large-scale studies conducted so far (e.g., Ferrand et al., 2010, 2017; New et al., 2006), hence allowing a direct comparison of the results.
3. Note, that if the standard orthographic (ON; Coltheart et al., 1977) and phonological (PN; Luce & Pisoni, 1998) neighbourhood measures were used instead, larger word length effects were observed in both tasks. Specifically, in the latency data from NAM, the percentage of unique variance increased to 8% for ON,  $F(20, 1910) = 176.98$ ,  $p < .001$ , and to 8.3% for PN,  $F(20, 1910) = 177.993$ ,  $p < .001$ , and in the latency data from LDT, to 4.2% both for ON,  $F(20, 1910) = 70.316$ ,  $p < .001$ , and PN,  $F(20, 1910) = 70.389$ ,  $p < .001$  neighbourhood measures. In the accuracy data from NAM, the percentage of unique variance increased to 0.8% both for ON,  $F(20, 1910) = 3.779$ ,  $p < .001$ , and PN,  $F(20, 1910) = 3.692$ ,  $p < .001$ , although in the accuracy data from LDT it remained the same (5.2%) for PN,  $F(20, 1910) = 20.545$ ,  $p < .001$ , and decreased slightly (4.7%) for ON,  $F(20, 1910) = 20.566$ ,  $p < .001$ .
4. Although Ernestus and Cutler (2015), Goh et al. (2016) and Ferrand et al. (2017) used orthographic and phonological uniqueness point measures in their megastudies,

here we only considered the orthographic uniqueness point measure since phonological uniqueness point measures are not available from the P-PAL database (see Soares, Iriarte, et al., 2018).

## Acknowledgements

This study was conducted at Psychology Research Centre (PSI/01662), University of Minho, and 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 COMPETE2020 under the PT2020 Partnership Agreement (POCI-01-0145-FEDER-007653). It is also part of the research project "Procura Palavras (P-Pal): A software program for deriving objective and subjective psycholinguistic indices for European Portuguese words" (PTDC/PSI-PCO/104679/2008).

## Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

Ana P. Pinheiro  <http://orcid.org/0000-0002-7981-3682>

## References

- 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*(9), 814–823. doi:10.1111/j.1467-9280.2006.01787.x
- Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*(4), 439–461. doi:10.3758/BF03214334
- Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language, 55*(2), 290–313. doi:10.1016/j.jml.2006.03.008
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*(2), 283–316. doi:10.1037/0096-3445.133.2.283
- Balota, D. A., Ferraro, F. R., & Connor, L. T. (1991). On the early influence of meaning in word recognition: A review of the literature. In P. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 187–222). Hillsdale: Erlbaum.
- Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition, 29*(4), 639–647. doi:10.3758/BF03200465
- Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition* (pp. 90–115). Hove: Psychology Press.
- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459. doi:10.3758/BF03193014
- Barca, L., Burani, C., & Arduino, L. S. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, & Computers, 34*(3), 424–434. doi:10.3758/bf03195471
- Bonin, P., Méot, A., & Bugajska, A. (2018). Concreteness norms for 1,659 French words: Relationships with other psycholinguistic variables and word recognition times. *Behavior Research Methods, 1*–22. doi:10.3758/s13428-018-1014-y
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*(1), 49–59. doi:10.1016/0005-7916(94)90063-9
- Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Gainesville, FL: Center for Research in Psychophysiology, University of Florida.
- 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*(5), 412–424. doi: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*(2), 422–430. doi: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*(4), 977–990. doi:10.3758/BRM.41.4.977
- Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch lexicon project 2. *Journal of Experimental Psychology: Human Perception and Performance, 42*(3), 441–458. doi:10.1037/xhp0000159
- Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica, 104*(2), 215–226. doi:10.1016/S0001-6918(00)00021-4
- Burani, C., Arduino, L. S., & Barca, L. (2007). Frequency, not age of acquisition, affects Italian word naming. *European Journal of Cognitive Psychology, 19*(6), 828–866. doi:10.1080/09541440600847946
- Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE, 5*(6), e10729. doi:10.1371/journal.pone.0010729
- Cameirão, M. L., & Vicente, S. G. (2010). Age-of-acquisition norms for a set of 1,749 Portuguese words. *Behavior Research Methods, 42*(2), 474–480. doi:10.3758/BRM.42.2.474
- Campos, A. D., Oliveira, H. M., & Soares, A. P. (2018). The role of syllables in intermediate-depth stress-timed languages: Masked priming evidence in European Portuguese. *Reading and Writing, 31*(5), 1209–1229. doi:10.1007/s11145-018-9835-8
- Campos, A. D., Soares, A. P., & Oliveira, H. M. (2018). Syllable effects in beginning and intermediate European-Portuguese readers: Evidence from a sandwich masked go/no-go lexical decision task. Manuscript under review for publication.
- Carroll, J. B., & White, M. N. (1973). Age-of-acquisition norms for 220 picturable nouns. *Journal of Verbal Learning and Verbal Behavior, 12*(5), 563–576.

- Chateau, D., & Jared, D. (2003). Spelling-sound consistency effects in disyllabic word naming. *Journal of Memory and Language*, 48(2), 255–280. doi:10.1016/S0749-596X(02)00521-1
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- 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.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108(1), 204–256. doi:10.1037/0033-295x.108.1.204
- Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology*, 60(8), 1072–1082. doi:10.1080/17470210701315467
- Cortese, M. J., & Schock, J. (2013). Imageability and age of acquisition effects in disyllabic word recognition. *Quarterly Journal of Experimental Psychology*, 66(5), 946–972. doi:10.1080/17470218.2012.722660
- Cuetos, F., & Barbón, A. (2006). Word naming in Spanish. *European Journal of Cognitive Psychology*, 18(3), 415–436. doi:10.1080/13594320500165896
- Cuetos, F., Glez-Nosti, M., Barbon, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, 32(2), 133–143.
- Davies, R., Barbón, A., & Cuetos, F. (2013). Lexical and semantic age-of-acquisition effects on word naming in Spanish. *Memory & Cognition*, 41(2), 297–311. doi:10.3758/s13421-012-0263-8
- Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *Quarterly Journal of Experimental Psychology*, 68(8), 1469–1488. doi:10.1080/17470218.2014.984730
- Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, 125(6), 777–799. doi:10.1037/0033-2909.125.6.777
- Fernandes, S., Ventura, P., Querido, L., & Morais, J. (2008). Reading and spelling acquisition in European Portuguese: A preliminary study. *Reading and Writing*, 21(8), 805–821. doi:10.1007/s11145-007-9093-7
- Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., ... Pallier, C. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: Evidence from Chronolex. *Frontiers in Psychology*, 2, 306. doi:10.3389/fpsyg.2011.00306
- Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., ... Grainger, J. (2017). MEGALEX: A megastudy of visual and auditory word recognition. *Behavior Research Methods*, 50(3), 1285–1307. doi:10.3758/s13428-017-0943-1
- Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., ... Pallier, C. (2010). The French lexicon project: Lexical decision data for 38,840 French words and 38,840 pseudo-words. *Behavior Research Methods*, 42(2), 488–496. doi:10.3758/brm.42.2.488
- Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, 28(7), 1109–1115. doi:10.3758/bf03211812
- Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35(1), 116–124. doi:10.3758/bf03195503
- Frost, R. (1998). Toward a strong phonological theory of visual word recognition: True issues and false trails. *Psychological Bulletin*, 123(1), 71–99. doi:10.1037/0033-2909.123.1.71
- Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113(2), 256–281. doi:10.1037/0096-3445.113.2.256
- Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, 115(1), 43–67. doi:10.1016/j.actpsy.2003.11.002
- Gimenes, M., Brysbaert, M., & New, B. (2016). The processing of singular and plural nouns in English, French, and Dutch: New insights from megastudies. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 70(4), 316–324. doi:10.1037/cep0000074
- Goh, W. D., Yap, M. J., Lau, M. C., Ng, M. M., & Tan, L. C. (2016). Semantic richness effects in spoken word recognition: A lexical decision and semantic categorization megastudy. *Frontiers in Psychology*, 7, 976. doi:10.3389/fpsyg.2016.00976
- González-Nosti, M., Barbón, A., Rodríguez-Ferreiro, J., & Cuetos, F. (2014). Effects of the psycholinguistic variables on the lexical decision task in Spanish: A study with 2,765 words. *Behavior Research Methods*, 46(2), 517–525. doi:10.3758/s13428-013-0383-5
- Goswami, U., & Ziegler, J. C. (2006). A developmental perspective on the neural code for written words. *Trends in Cognitive Sciences*, 10(4), 142–143. doi:10.1016/j.tics.2006.02.006
- Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (3rd ed.). New York: Macmillan.
- Izura, C., Wright, V. C., & Fouquet, N. (2014). Hemispheric asymmetries in word recognition as revealed by the orthographic uniqueness point effect. *Frontiers in Psychology*, 5, 244. doi:10.3389/fpsyg.2014.00244
- Kessler, B., Treiman, R., & Mullennix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language*, 47(1), 145–171. doi:10.1006/jmla.2001.2835
- Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Quarterly Journal of Experimental Psychology*, 68(8), 1457–1468. doi:10.1080/17470218.2015.1051065
- Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, 1, 174. doi:10.3389/fpsyg.2010.00174
- Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British lexicon project: Lexical decision data for 28,730 mono-syllabic and disyllabic English words. *Behavior Research Methods*, 44(1), 287–304. doi:10.3758/s13428-011-0118-4
- Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words:

- Why emotion matters. *Journal of Experimental Psychology: General*, 140(1), 14–34. doi:10.1037/a0021446
- Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion. *Quarterly Journal of Experimental Psychology*, 68(8), 1693–1710. doi:10.1080/17470218.2014.989865
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978–990. doi:10.3758/s13428-012-0210-4
- 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(2), 376–381. doi:10.1037/0096-1523.25.2.376
- Lima, C. F., & Castro, S. L. (2010). Reading strategies in orthographies of intermediate depth are flexible: Modulation of length effects in Portuguese. *European Journal of Cognitive Psychology*, 22(2), 190–215. doi:10.1080/09541440902750145
- Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics*, 39(3), 155–158. doi:10.3758/bf03212485
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19(1), 1–36. doi:10.1097/00003446-199802000-00001
- New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English lexicon project. *Psychonomic Bulletin & Review*, 13(1), 45–52. doi:10.3758/bf03193811
- Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76, 1–25. doi:10.1037/h0025327
- 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), 1–17. doi:10.1080/17470218.2015.1105268
- Perea, M. (2015). Neighborhood effects in visual-word recognition and reading. In A. Pollatsek & R. Treiman (Eds.), *The Oxford handbook on reading* (pp. 76–87). New York: Oxford University Press.
- Perea, M., Comesaña, M., & Soares, A. P. (2012). Does the advantage of the upper part of words occur at the lexical level? *Memory & Cognition*, 40(8), 1257–1265. doi:10.3758/s13421-012-0219-z
- 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(1), 37–44. doi:10.1016/j.jecp.2012.10.014
- Perea, M., Winkler, H., & Gomez, P. (2018). How orthographic-specific characteristics shape letter position coding: The case of Thai script. *Psychonomic Bulletin & Review*, 25(1), 416–422. doi:10.3758/s13423-017-1279-7
- Perfetti, C. A., Zhang, S., & Berent, I. (1992). Reading in English and Chinese: Evidence for a ‘universal’ phonological principle. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 227–248). Amsterdam: Elsevier North-Holland.
- Pinheiro, A. P., Dias, M., Pedrosa, J., & Soares, A. P. (2017). Minho affective sentences (MAS): Probing the role of sex, mood and empathy in affective ratings of verbal stimuli. *Behavior Research Methods*, 49(2), 698–716. doi:10.3758/s13428-016-0726-0
- Protopapas, A. (2007). Check vocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods*, 39(4), 859–862. doi:10.3758/bf03192979
- Pureza, R., Soares, A. P., & Comesaña, M. (2016). Cognate status, syllable position and word length on bilingual tip-of-the-tongue states induction and resolution. *Bilingualism: Language and Cognition*, 19(3), 533–549. doi:10.1017/s1366728915000206
- Radeau, M., Mousty, P., & Bertelson, P. (1989). The effect of the uniqueness point in spoken-word recognition. *Psychological Research*, 51(3), 123–128. doi:10.1007/bf00309307
- Rey, A., & Courrieu, P. (2010). Accounting for item variance in large-scale databases. *Frontiers in Psychology*, 1, 200. doi:10.3389/fpsyg.2010.00200
- Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94(2), 143–174. doi:10.1348/000712603321661859
- Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., & Frade, C. S. (2012). The adaptation of the affective norms for English words (ANEW) for European Portuguese. *Behavior Research Methods*, 44(1), 256–269. doi:10.3758/s13428-011-0131-7
- Soares, A. P., Costa, A. S., Machado, J., Comesaña, M., & Oliveira, H. M. (2017). The Minho word pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. *Behavior Research Methods*, 49(3), 1065–1081. doi:10.3758/s13428-016-0767-4
- Soares, A. P., Iriarte, A., Almeida, J. J., Simões, Á, 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), 110–123. doi:10.1590/S0102-79722014000100013
- Soares, A. P., Iriarte, Á, Almeida, J. J., Simões, A., Costa, A., Machado, J., ... Perea, M. (2018). Procura-PALavras (P-PAL): A web-based interface for a new European Portuguese lexical database. *Behavior Research Methods*, 50(4), 1461–1481. doi:10.3758/s13428-018-1058-z
- Soares, A. P., Lages, A., Oliveira, H. M., & Hernández, J. (2018). The mirror reflects more for ‘d’ than for ‘b’: Right-asymmetry bias on the visual recognition of words containing reversal letters. Manuscript under review for publication.
- Soares, A. P., Machado, J., Costa, Á, Iriarte, A., Simões, A., Almeida, J. J. ... Perea, M. (2015). On the advantages of frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. *The Quarterly Journal of Experimental Psychology*, 68(4), 680–696. doi:10.1080/17470218.2014.964271
- Soares, A. P., Medeiros, J. C., Simões, A., Machado, J., Costa, A., Iriarte, Á, ... Comesaña, M. (2014). ESCOLEX: A grade-level lexical database from European Portuguese elementary to middle school textbooks. *Behavior Research Methods*, 46(1), 240–253. doi:10.3758/s13428-013-0350-1
- Soares, A. P., Oliveira, H. M., Comesaña, M., & Costa, A. S. (2018). Lexico-syntactic interactions in the resolution of relative clause ambiguities in a second language (L2): The role cognate status and L2 proficiency. *Psicológica Journal*, 39(2), 164–197. doi:10.2478/psicolj-2018-0008

- Soares, A. P., Oliveira, H. M., Ferreira, M., Comesaña, M., Macedo, A. F., Ferré, P., ... Fraga, I. (2018). Lexico-syntactic interactions during the processing of temporally ambiguous L2 relative clauses: An eye-tracking study with intermediate and advanced Portuguese-English bilinguals. Manuscript submitted for publication.
- Soares, A. P., Perea, M., & Comesaña, M. (2014). Tracking the emergence of the consonant bias in visual-word recognition: Evidence with developing readers. *PLoS ONE*, *9*(2), e88580. doi:10.1371/journal.pone.0088580
- Soares, A. P., Pinheiro, A. P., Costa, A., Frade, C. S., Comesaña, M., & Puresa, R. (2015). Adaptation of the international affective picture system (IAPS) for European Portuguese. *Behavior Research Methods*, *47*, 1159–1177. doi:10.3758/s13428-014-0535-2
- Soares, A. P., Pinheiro, A. P., Costa, A., Frade, S., Comesaña, M., & Puresa, R. (2013). Affective auditory stimuli: Adaptation of the international affective digitized sounds (IADS-2) for European Portuguese. *Behavior Research Methods*, *45*(4), 1168–1181. doi:10.3758/s13428-012-0310-1
- Spieler, D. H., & Balota, D. A. (2000). Factors influencing word naming in younger and older adults. *Psychology and Aging*, *15*(2), 225–231. doi:10.1037/0882-7974.15.2.225
- Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78. doi:10.1207/s15516709cog2901\_3
- Sze, W. P., Rickard Liow, S. J., & Yap, M. J. (2014). The Chinese lexicon project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, *46*(1), 263–273. doi:10.3758/s13428-013-0355-9
- Sze, W. P., Yap, M. J., & Rickard Liow, S. J. (2015). The role of lexical variables in the visual recognition of Chinese characters: A megastudy analysis. *Quarterly Journal of Experimental Psychology*, *68*(8), 1541–1570. doi:10.1080/17470218.2014.985234
- Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General*, *124*(2), 107–136. doi:10.1037/0096-3445.124.2.107
- Tsang, Y.-K., Huang, J., Lui, M., Xue, M., Chan, Y.-W. F., Wang, S., & Chen, H.-C. (2017). MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior Research Methods*, doi:10.3758/s13428-017-0944-0
- Tse, C.-S., Yap, M. J., Chan, Y.-L., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese lexicon project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*, *49*(4), 1503–1519. doi:10.3758/s13428-016-0810-5
- 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*(6), 1176–1190. doi:10.1080/17470218.2013.850521
- Vasconcelos, M., Dias, M., Soares, A. P., & Pinheiro, A. P. (2017). What is the melody of that voice? Probing unbiased recognition accuracy with the Montreal affective voices. *Journal of Nonverbal Behavior*, *41*(3), 239–267. doi:10.1007/s10919-017-0253-4
- Wilson, M. A., Cuetos, F., Davies, R., & Burani, C. (2013). Revisiting age-of-acquisition effects in Spanish visual word recognition: The role of item imageability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1842–1859. doi:10.1037/a0033090
- Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*(4), 502–529. doi:10.1016/j.jml.2009.02.001
- Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay lexicon project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, *42*(4), 992–1003. doi:10.3758/brm.42.4.992
- Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, *18*(4), 742–750. doi:10.3758/s13423-011-0092-y
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. doi:10.3758/pbr.15.5.971
- Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47*(1), 1–29. doi:10.1006/jmla.2001.2834
- Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*(1), 3–29. doi:10.1037/0033-2909.131.1.3
- Ziegler, J. C., Tan, L. H., Perry, C., & Montant, M. (2000). Phonology matters: The phonological frequency effect in written Chinese. *Psychological Science*, *11*(3), 234–238. doi:10.1111/1467-9280.00247