Behavior Research Methods (2023) 55:3699–3725
https://doi.org/10.3758/s13428-022-01985-3

PHOR-in-One: A multilingual lexical database with PHonological, ORthographic and PHonographic word similarity estimates in four languages

Ana Santos Costa[1] · Montserrat Comesaña[1,2] · Ana Paula Soares[1]

Accepted: 20 September 2022 / Published online: 7 November 2022

Abstract

A large body of research seeking to explore how form affects lexical processing in bilinguals has suggested that orthographically similar translations (e.g., English-Portuguese "paper-*papel*") are responded to more quickly and accurately than words with little to no overlap (e.g., English-Portuguese "house-*casa*"). One of the most prominent algorithms to estimate orthographic similarity, the normalized Levenshtein distance (NLD), returns an index of the proportion of identical characters of two strings, and is an efficient and invaluable tool for the selection, manipulation, and control of verbal stimuli. Notwithstanding its many advantages for second-language research, the absence of a comparable measure for phonology has resulted in the adoption of different strategies to assess the degree of interlanguage phonological similarity across the literature, with profound implications for the interpretation of results on the relative role of orthographic and phonological similarity in bilingual lexical access. In the present work, we introduce PHOR-in-One, a multilingual lexical database with a set of phonological and orthographic NLD estimates for 6160 translation equivalents in American and British English, European Portuguese, German and Spanish in a total of 30,800 words. We also propose a new measure of phonographic NLD, a pooled index of orthographic and phonological similarity, particularly useful for researchers interested in controlling for and/or manipulating both estimates at once. PHOR-in-One includes a comprehensive characterization of its lexical entries, namely Part-of-Speech-dependent and independent frequency counts, number of letters and phonemes, and phonetic transcription. PHOR-in-One is thus a valuable tool to support bilingual and multilingual research.

Keywords NLD · Orthographic similarity · Phonological similarity · Phonographic similarity · Cognates · Levenshtein distance · PHOR-in-One

The efficiency with which bilinguals make language-related decisions in the face of similarity is a truly remarkable feat. When visually confronted with nearly identical translation equivalents, such as English-Portuguese "paper-*papel*", speakers can quickly assign them to different languages, retrieve correct pronunciations from different phonological systems and ascribe meaning. Interlanguage form similarity has been widely studied in both experimental (e.g., Brenders et al., 2011; Christoffels et al., 2006; Comesaña et al., 2012, 2015; Costa et al., 2000, 2005; de Groot & Nas, 1991; Dijkstra et al., 1999, 2010; Lemhöfer & Dijkstra, 2004; Soares et al., 2018b, 2019b; van Hell & Dijkstra, 2002; Voga & Grainger, 2007) and real-life (e.g., Cunningham & Graham, 2000; Holmes & Ramos, 1993; Peters & Webb, 2018) settings, and the bulk of research has suggested that there is a processing advantage when translations formally resemble one another. For instance, when asked to recognize (e.g., Dijkstra et al., 2010; Ferré et al., 2017; Lemhöfer et al., 2004; Voga & Grainger, 2007) or name (e.g., Broersma et al., 2016; Costa et al., 2000, 2005; Hoshino & Kroll, 2008) words in a foreign language, bilinguals typically provide faster and more accurate responses to cognates, i.e., translation equivalents with similar (e.g., English-Portuguese "theory-*teoria*") or identical (e.g., "animal-*animal*") form, than to noncognates (words that share

✉ Ana Santos Costa
anamariacscosta@gmail.com

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

[2] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Madrid, Spain

their meaning but not their form, e.g., English-Portuguese "house-*casa*"). They have also been shown to learn (e.g., de Groot & Keijzer, 2000; Lotto & de Groot, 1998; Peters & Webb, 2018; Valente et al., 2017) and remember (e.g., de Groot & Keijzer, 2000) cognates more effectively than noncognates (see however Arana et al., 2022 and Comesaña et al., 2012, 2015, for reversed or null effects as a function of stimulus list composition; see also Pureza et al., 2016 for poorer resolution of Tip-of-the-Tongue states in cognates vs. noncognates).

To examine how similarity affects processing in bilinguals and multilinguals, researchers typically retrieve the most accurate translation for the words in their stimulus sets, and subsequently assess the extent to which their forms overlap, either by collecting subjective ratings (i.e., asking bilinguals to provide personal similarity judgements via Likert-type scales), or using computational algorithms, such as Van Orden's orthographic similarity (Van Orden, 1987), which indicates the proportion of identical characters shared by two words, and the Levenshtein distance (LD; Levenshtein, 1966), which returns the minimum number of insertions, deletions and/or substitutions required to transform one string into another. Recently, the normalized LD (NLD; Schepens et al., 2012), estimated by dividing the LD of two words by the number of letters in the longest string and subtracting the result from one, has become prevalent in the literature for a number of reasons. First, varying on a continuum between 0 (no similarity, e.g., English-Portuguese "sky-*céu*") and 1 (exact match, e.g., "banana-*banana*"), the result is easily interpretable and allows for an objective characterization of translation equivalents as identical cognates, non-identical cognates or noncognates, with a threshold of .50 typically considered for cognate inclusion (Schepens et al., 2012). Second, it is sensitive to word length, and thus character additions and deletions have a smaller impact in the final NLD in longer than shorter words (for instance, in the English-European Portuguese translation pairs "air-*ar*" and "intellectual-*intelectual*", even though only one character is deleted in each pair to transform one word into the other, the NLD is .67 in the former and .92 in the latter). Third, it is independent of word and language order, i.e., the orthographic NLD for the translation pair "house-*casa*" is .20 regardless of whether the input is "house-*casa*" or "*casa*-house". Unlike Van Orden's orthographic similarity, this is a key feature in this algorithm, since it satisfies the commutative property of similarity (Guasch et al., 2013). Lastly, web-accessible tools (e.g., Guasch et al., 2013) that allow for the submission of large lists of translation equivalents and instantly return their NLD have also contributed to its growing popularity.

The efficiency of the NLD has turned it into the gold standard for the analysis of orthographic overlap, but the shortage of equivalent indices for phonology has led to a range of approaches in the assessment of phonological similarity. Even though a few existing lexical databases already return some sort of phonological information for multiple languages, e.g., Celex (English, German, and Dutch; Baayen et al., 1995), CLEARPOND (English, Dutch, German, Spanish, and French; Marian et al., 2012), or WordGen (English, German, Dutch, and French; Duyck et al., 2004), only CLEARPOND performs cross-language comparisons (Marian, 2017), but it focuses solely on neighborhood analyses, providing information on the words that can be formed in one language by changing one phoneme in a word of another language, e.g., the English word "height" [haɪt] and its German phonological neighbors "*hat*" [hat] (he/she/it has; deletion of [ɪ]), "*heiß*" [haɪs] (hot; replacement of [t] with [s]) and "*seit*" [zaɪt], (since; replacement of [h] with [z]). Consequently, selecting and characterizing translation equivalents in terms of how phonologically similar they are from a reliable, centralized tool is still virtually impossible, leading to the adoption of different strategies, from time-consuming collections of subjective ratings (e.g., Brenders et al., 2011; Dijkstra et al., 1999, 2010; Hoshino & Kroll, 2008; Poort & Rodd, 2019; Post da Silveira & van Leussen, 2015; Schwartz et al., 2007) to manual analyses of the proportion of identical phonetic syllables or segments in the source and target strings (e.g., Blumenfeld & Marian, 2005; Comesaña et al., 2012, 2015; Costa et al., 2000; Valente et al., 2017; Voga & Grainger, 2007). In addition, the difficulty in promptly accessing reliable estimates of phonological similarity for sufficiently large stimulus sets has widened the gap between the number of studies exploring the role of orthographic and phonological similarity in second-language processing, as already acknowledged in the literature (see Comesaña et al., 2015 and Dijkstra et al., 1999 for an overview; see also Dijkstra et al., 2010), and has critical implications for the tenets of bilingual computational models (e.g., Dijkstra et al., 2019; Dijkstra & van Heuven, 2002; van Heuven et al., 1998), most of which are rooted on orthographic similarity alone (Comesaña et al., 2015).

More recently, some bilingual and multilingual lexical databases offering estimates of interlanguage phonological similarity were put forth in the literature, but with considerable limitations nonetheless. Poort and Rodd (2019) advanced an English-Dutch stimulus set containing translation equivalents with varying degrees of form similarity, and interlingual homographs (i.e., words with similar form but different meaning in two languages, e.g., "angel-*angel*" – heavenly being vs. sting of bee or wasp), and Post da Silveira and van Leussen (2015) proposed a bilingual lexicon comprising equisyllabic cognates and noncognates for Brazilian Portuguese-American English. Despite the benefits of these resources for second-language research, only very small stimulus sets (Poort & Rodd, 2019: 58 identical cognates, 76 non-identical cognates, 78 noncognates, and 72 interlingual homographs; Post da Silveira & van Leussen, 2015: 64 cognates and 40 noncognates) were included in a single language combination in each database. In addition, while the phonological similarity estimates in the Poort and Rodd study were based on subjective ratings, Post da Silveira and van Leussen used the classical NLD, which does not take phoneme similarities into account, and hence their objective indices of phonological similarity may be slightly underestimated (for instance, in the American English-Brazilian Portuguese translation pair "minister-*ministro*" [ˈmɪnɪstəɹ-mi'nistɾu], substituting the nearly identical phonemes /ɪ/ with /i/ and /ɹ/ with /ɾ/ is as costly as any other consonant-vowel substitution). To this end, a more fine-tuned, NLD-based phonological similarity index was proposed by Schepens (Schepens, 2010; see also Schepens et al., 2013) in a lexical database for multiple languages that compared the phonetic transcriptions of two translation equivalents, and applied a modified version of the LD that introduced the degree of similarity between the source and target phonemes as substitution costs. However, the materials only comprised high-frequency cognates, and thus researchers interested in manipulating different frequency values or form similarity degrees (e.g., low and medium-frequency words and noncognates) cannot retrieve such stimuli from this database.

Another issue in reference to the assessment of interlanguage form similarity concerns the absence of an index of phonographic similarity in the literature, i.e., a method that combines the degree of orthographic and phonological overlap of two translation equivalents in a single measure. In effect, similar measures of phonographic similarity within languages have been examined with monolingual populations, and suggested that phonographic effects outperform orthographic and/or phonological measures taken separately. For instance, to test whether onset-nucleus-coda subsyllabic components mediate visual word recognition, Nuerk and collaborators (2000) introduced a measure of subcomponent frequency (SCF) for phonographic sublexical units, i.e., orthographic subsets that are phonology-dependent (e.g., the bigram <or> in "horse" [hɔːs] and "morse" [mɔːs], but not in "worse" [wɜːs]). Results from a lexical decision task revealed that the phonographic SCF facilitated visual word recognition, and that bigram frequency (a purely orthographic measure) did not produce an effect when SCF was controlled for. Research looking into within-language neighborhoods has also shown that the phonographic *N* (the number of words of equal length in letters and phonemes that can be generated by a single letter and phoneme substitution, addition or deletion; Peereman & Content, 1997; Siew & Vitevitch, 2019) produces a processing advantage for words that have more rather than fewer phonographic neighbors, and that it is a more significant predictor of subjects' performances than the orthographic and phonological *N* individually (see Adelman & Brown, 2007, and Siew & Vitevitch, 2019 for overviews). Although an interlanguage phonographic similarity measure has never been investigated with bilingual or multilingual populations, a number of studies using different types of stimuli have shown the close interdependence of orthographic and phonological interactions. For instance, in cognate recognition and naming, performances become altered under the influence of phonologically similar words in another language, even when using language pairs with different scripts, e.g., Greek-Spanish (Dimitropoulou et al., 2011), Greek-French (Voga & Grainger, 2007), Japanese-English (Ando et al., 2014) and Hebrew-English (Gollan et al., 1997). In a slightly different line of research, using interlingual pseudo-homophones, i.e., pseudowords that sound identical to real words (e.g., "*tauw*" is a pseudo-homophone of the Dutch word "*touw*", meaning "rope") as cross-language primes to their translation equivalents in English ("rope"), Duyck (2005) found differences in performances when English targets were preceded by pseudo-homophone primes compared to their graphemic controls. The fact that these phonological similarity effects were observed during silent reading, often with masked priming (a useful technique to investigate the activation of orthographic, phonological, and semantic codes that influence the early stages of visual word recognition) or pseudo-homophones (which do not involve activation of semantic representations), reinforces the assumption that the presentation of orthographic input necessarily activates a fast and automatic phonological representation that does not require lexical access (Dimitropoulou et al., 2011). These findings on the strong reciprocity of orthographic and phonological activation at pre-lexical stages of bilingual processing (see also Clifton, 2015 and Schotter et al., 2012 for a review of studies investigating the influence of phonological similarity during parafoveal processing in reading with monolinguals and bilinguals) lend strong support to the assumption that an interlanguage phonographic similarity measure could be a more solid predictor of bilinguals' performances than the orthographic and phonological NLDs in isolation.

To address these limitations, here we introduce PHOR-in-One, a fully integrated multilingual lexical database containing 6160 translation equivalents in English – both American and British varieties – European Portuguese, German, and Spanish, and an array of interlanguage form similarity measures, including the classical orthographic NLD, an adapted phonological NLD, and a phonographic NLD as an estimate of the degree of overall form similarity of two translation equivalents. Based on an adaptation of Schepens' (2010) algorithm, the phonological NLD considers the articulatory, acoustic, and perceptive features of the phonemes as substitution costs. For instance, to determine the phonological NLD of the English-Portuguese pair "house-*casa*", the algorithm takes their phonetic transcriptions [ʻhaʊs-ʻkazɐ], identifies the relative positions of the phonemes to be replaced ([k] with [h], [a] with [aʊ], and [s] with [z]) from the

International Phonetic Alphabet (IPA; International Phonetic Association, 1999) feature space, computes the Euclidean distance between them, and finally adds the cost of inserting [ɐ]. It should be noted that substantial modifications to the original algorithm (Schepens, 2010) were introduced for our computation, as discussed ahead. The total transformation cost is subsequently normalized, allowing for the characterization of translation pairs as phonologically identical (e.g., English-German “fish-*Fisch* [‘fɪʃ-‘fɪʃ], phonological NLD = 1), phonologically similar (e.g., English-Portuguese “voice-*voz*” [‘vɔɪs-‘vɔʃ], phonological NLD = .90), and phonologically dissimilar (e.g., English-Portuguese “age-*idade*” [‘eɪdʒ- i‘ðaðɨ], phonological NLD = 0.25), all of which are included in our database. As for the phonographic NLD, it is computed by intersecting the orthographic and phonological NLDs of each translation pair (see ahead for details) and also ranges between 0 (entirely different orthography and phonology, e.g., American English-European Portuguese “butler-*mordomo*” [‘bʌtłəɹ̠-mɔɾ’ðomu]) and 1 (orthographically and phonologically identical words, e.g., English-German “test-*Test*” [‘tɛst-‘tɛst]). To our knowledge, an interlanguage phonographic similarity measure has never been advanced in the literature, but examining the overlapping areas of the phonological and orthographic layers may offer valuable insights on the interaction of phonology and orthography in various language processes (Siew & Vitevitch, 2019). Overall, the three cross-language form similarity estimates in PHOR-in-One will allow experimenters to manipulate different degrees of orthographic, phonological, and phonographic overlap using continuous variables, rather than relying on an arbitrary threshold to define cognates (Tainturier, 2019). The orthographic and phonological NLD will be particularly useful to select stimuli with contrasting degrees of orthographic and phonological overlap (O+P-, O-P+), while the phonographic NLD can be used to select stimuli with high (O+P+) or low (O-P-) orthographic and phonological overlap. Although future studies should compare the three estimates, and explore how well they can capture subjects' performances, they will contribute to test recent accounts (e.g., Iniesta et al., 2021) of the organization of phonological and orthographic interactions in visual and auditory processing in bilinguals.

Aside from the interlanguage form similarity estimates, PHOR-in-One comprises a range of features to ensure a comprehensive characterization of its stimuli within and across languages. Relevant linguistic information, such as number of letters and phonemes, phonetic transcription, and a number of frequency indices are provided for the full multilingual lexicon in a total of 30,800 words. Its easy-to-use format enables the automatic retrieval of words and their translation equivalents in each language, based on the application of specific linguist criteria (e.g., Part-of-Speech [PoS], NLD interval and per-million-word frequency). In addition, PHOR-in-One contains languages with different opacities (transparent: German and Spanish; intermediate: European Portuguese; opaque: English), timings (stress-timed: English, European Portuguese and German; syllable-timed: Spanish; see Nespor et al., 2011 for an overview of rhythm and timing, and also Campos et al., 2018 for an example of how timing can affect the role of sublexical units in processing) and families (Romance languages: European Portuguese and Spanish; Germanic languages: English and German). It also includes different word types, such as simple and compound words, multi-word expressions, identical and non-identical cognates and noncognates from all ranges of frequency and interlanguage form similarity. Distributed along a continuum of orthographic, phonological, and phonographic similarity, the different types of stimuli in PHOR-in-One will further encourage the development of research to examine how language transparency affects processing, and the circumstances under which bilinguals rely on grapheme-to-phoneme mappings or on more direct access to whole-word representations across languages (Iniesta et al., 2021). PHOR-in-One is thus a useful research instrument, in that it delivers the most fundamental measures for the selection, control and manipulation of experimental multilingual materials.

Materials and methods

PHOR-in-One lexical database

Entry compilation and translation procedures

The PHOR-in-One lexicon originated from two existing stimulus sets used previously as experimental materials at our lab, one containing 5048 European Portuguese, English and Spanish translation equivalents, and another one with 1779 translation equivalents in European Portuguese, English, and German. The two sets comprised words with an array of interesting features for research, including short, medium, and long words, as well as cognates and noncognates with low, medium, and high lexical frequency. Intersection of the two sets revealed the existence of 871 words in common, which, upon removal, originated an integrated lexicon of 5956 unique entries. Additional lexical entries were included for homonymous and polysemic words if they originated more than one orthographic form in another language. For instance, the English word “hug” generates different words in German, European Portuguese and Spanish, depending on whether the grammatical category is a noun (“*Umarmung*”, “*abraço*”, “*abrazo*”) or a verb (“*umarmen*”, “*abraçar*”, “*abrazar*”). To incorporate different forms in the three languages, two separate lexical entries were created, where the English word “hug” is duplicated. In the same vein, the European Portuguese noun "*canto*" translates as the German words "*Eckball*" (football corner), "*Kante*" or "*Ecke*" (corner of a room or table), and "*Gesang*" (singing), and hence four separate lexical entries were created to accommodate four different words in German, although the same word "*canto*" is displayed for European Portuguese. Unavailable German and Spanish translations in each list were added automatically. An American English lexicon was created from the British English words by adapting the spellings. The two varieties were included on account of their differences in terms of: i) terminology (e.g., British English "chemist, lift" vs. American English "drugstore, elevator"); ii) spelling (e.g., British English "characterise, honour, centre" vs. American English "characterize, honor, center"); iii) frequency of use (e.g., the word "analysis" has a raw frequency of 8456 occurrences in the original American English corpus, and 1101 in the British English corpus; conversely, the word "back" appears more frequently in British than American English, with 22,071 and 17,570 raw counts, respectively); and iv) pronunciation (approximately 70% of the lexical entries in PHOR-in-One present different phonetic transcriptions for American and British English). An expert in the four languages subsequently conducted a comprehensive review of the translations, applying corrections where needed. Finally, two native speakers of German and Spanish with different second-language combinations reviewed the translations.

The resulting multilingual lexicon includes 6160 lexical entries, each containing a wordform in American and British English, German, European Portuguese and Spanish, all fully aligned across languages, in a total of 30,800 words. Because of their features, the words in each language are particularly useful to cover a broad range of research requirements for the manipulation and/or control of verbal stimuli. The five lexica include a) words with varying lengths, ranging between two and 22 letters (American and British English: *min* = 2 ["go"]; *max* = 16 ["misunderstanding"]; German: *min* = 2 ["*Ei*"; egg]; *max* = 22 ["*Erziehungsberechtigter*"; guardian]; European Portuguese: *min* = 2 ["*pó*"; dust]; *max* = 16 ["*congestionamento*"; jam]; Spanish: *min* = 2 ["*té*"; tea]; *max* = 15 ["*existencialismo*"; existentialism]); b) low, medium and high-frequency words, ranging between 0.01 and 7903.62 occurrences per million words (pmw; American English: *min* = 0.02 ["adequacy"]; *max* = 6161.41 ["have"]; British English: *min* = 0.01 ["artifact"]; *max* = 7903.62 ["have"]; German: *min* = 0.03 ["*File*"; file] *max* = 4201.37 ["*haben*"; have]; European Portuguese: *min* = 0.01 ["*ermida*"; hermitage]; *max* = 5512.72 ["*bem*"; well]; Spanish: *min* = 0.02 ["*ánodo*"; anode]; *max* = 5804.59 ["*bien*"; well]; SUBTLEX pmw occurrences in each language); and c) words from five different grammatical categories, namely nouns, verbs, adjectives, adverbs and interjections (in addition to five compound grammatical categories, as detailed ahead). Moreover, all languages include simple (e.g., "house"), closed-compound (items containing at least two stems, Lieber, 2010; e.g., "notebook"), and hyphenated (e.g., "t-shirt") words (except for Spanish, which does not include hyphenated words). As an inherent consequence of the translation process, simple and closed-compound words in one language often originate multi-word expressions (items containing at least two words with unitary semantic or pragmatic function, Moon, 2015) in another language. Even though they are typically not included in analogous lexical databases, and make up for only a small portion of the PHOR-in-One lexica (American English: 0.73% of the lexicon; British English: 0.71%; German: 0.15%; European Portuguese: 0.08%; Spanish: 0.10%), we opted to preserve multi-word expressions (e.g., English-Portuguese "nut-*fruto seco*") including phrasal (e.g., "warm up") and reflexive (e.g., German "*sich erinnern*", to remember) verbs, as they contribute to promote lexical diversity, while opening a window of opportunity to explore whether there may be processing differences between them and simple or closed-compound words (see Arnon & Christiansen, 2017 and Titone & Libben, 2014 for an overview of the role of multi-word expressions in language learning abilities).

Spelling and pronunciation in each lexicon comply with the orthographic entries and phonetic transcriptions adopted in monolingual and multilingual dictionaries, namely the Dictionary of the Contemporary Portuguese Language (Casteleiro, 2001), the Dictionary of the Spanish Language (Real Academia Española, n.d.), the Duden Dictionary (Dudenredaktion, n.d.), and the Oxford English Dictionary (Oxford University Press, n.d.) for European Portuguese, Spanish, German, and English, respectively, which mirror the standard linguistic varieties of the corresponding languages. It is worth mentioning that Castilian Spanish was considered for the Spanish lexicon, and that the European Portuguese spelling reflects the norm before the 1990 Portuguese Language Orthographic Agreement, implemented in Portuguese-speaking countries in 2015, since there are no frequency norms available thus far in the literature for post-Agreement spelling. However, an additional column was created to accommodate the new spelling ($N = 115$; note that pre- and post-Agreement mismatches will not influence the phonological or phonographic similarity indices, since pronunciation has been maintained). Noun capitalization was preserved in German, as determined by convention, while other grammatical categories are lowercase. Some words are also capitalized in American and British English, namely those referring to nationalities and languages (e.g., "American"), months (e.g., "April"), holidays (e.g., "Christmas"), and some proper nouns (e.g., "Earth"; proper nouns were however generally excluded from PHOR-in-One, since they are typically not relevant for behavioral research).

Frequency and PoS assignment

Two kinds of wordform frequency information are offered in PHOR-in-One, namely monolingual printed corpus frequency, including spoken (e.g., conversations, interviews) and written (e.g., books, newspapers) records, and subtitle frequency. Although subtitle frequency accounts for more variance in lexical decision and naming tasks (see Brysbaert & New, 2009; Soares et al., 2014a, 2019a for reviews), some SUBTLEX studies lack relevant morphosyntactic information (e.g., SUBTLEX-ESP and SUBTLEX-DE do not include grammatical annotation), and as such, combined, the two resources offer a more comprehensive characterization of the words in the five lexica. Corpus frequency information was retrieved from the following preexisting lexical databases: the American National Corpus (ANC; Ide, 2009) for American English, Celex (Baayen et al., 1995) for British English, dlexDB (Heister et al., 2011) for German, ESPAL (Duchon et al., 2013) for Spanish, and P-PAL (Soares et al., 2018a) for European Portuguese. Subtitle frequencies in each language were retrieved from the corresponding preexisting SUBTLEX studies (American English: Brysbaert et al., 2012; British English: van Heuven et al., 2014; European Portuguese: Soares et al., 2014a; German: Brysbaert et al., 2011; Spanish: Cuetos et al., 2011). Absolute (raw counts), pmw and $\log_{10}$ (estimated by determining the base 10 logarithm of the absolute frequency counts + 1) wordform frequency norms are provided for each lexical entry for both printed and subtitle corpora. Certain frequency measures were unavailable in some of the original lexical databases (e.g., the ANC; Ide, 2009), and were thus calculated by us for cross-language comparability. For subtitle frequencies, Zipf norms (van Heuven et al., 2014) were also included, as long as they were available in the original SUBTLEX databases. Zipf frequency norms offer several advantages compared to other frequency indices, since they resemble a Likert scale and include the size of the corpus in the computation. The scale is intuitive to allow for a clear distinction between low- and high-frequency words (van Heuven et al., 2014), but unlike $\log_{10}$ frequencies offers a straightforward relationship with pmw frequency, with the values 1–3 indicating low-frequency words (frequencies of 1 per million words and lower) and the values 4–7 indicating high-frequency words (frequencies of 10 per million words and higher; Brysbaert & New, 2009; see also Soares et al., 2014a).

For corpus frequencies, two types of counts were included, namely PoS-independent (an index of the total number of times a wordform appears in the corpus, e.g., the word “act” appears 3354 times in the British English corpus overall), and PoS-dependent (an index of the number of times a wordform appears in the corpus with a specific grammatical category, e.g., the word “act” appears 2278 times as a noun and 1076 times as a verb in the British English corpus) frequency indices. As with other recent lexical databases (e.g., Duchon et al., 2013; Soares et al., 2018a), the two indices were included due to the existence of systematic differences in the processing of different grammatical categories (see Brysbaert et al., 2012 for a review), and hence some researchers may be more interested in intersecting frequency with PoS, rather than simply extracting overall frequency information.

The following five major grammatical categories were created: noun (75.07% of the full lexicon), verb (14.69%), adjective (9.40% of the lexicon), adverb (0.08%) and interjection (0.05%). Five additional compound categories, namely adjective-adverb (ADJ|ADV), adjective-noun (ADJ|N), adjective-noun-quantifier (ADJ|N|QUANT), adjective-noun-verb (ADJ|N|V), and adjective-verb (ADJ|V) were incorporated, since some words were originally annotated with complex PoS tags in the source corpora (approximately 0.71% of the full PHOR-in-One lexicon), and hence it was not possible to split their frequency into individual grammatical categories. The same PoS tag is true for the five languages in each lexical entry (note however that a language-specific PoS tag was added elsewhere for words whose PoS-dependent frequency information is indexed to a different grammatical category; see ahead for details). Data cleaning procedures were implemented during PoS and frequency compilation, which, in some cases, produced small frequency variations compared to the source lexical databases. Specifically, wordform frequencies originally annotated in the source corpora as non-word items, unknown/unclassified categories, web elements, dates and proper nouns were subtracted from the frequency of the corresponding lexical entries in PHOR-in-One. Subclass frequency counts, when available in the original lexical databases, were added to a single main grammatical class. For instance, the verb “accept” has an overall raw frequency of 1704 occurrences in PHOR-in-One, even though in Celex (Baayen et al., 1995) this value is split into four equal parts (426 occurrences in four separate lexical entries), each reflecting potential occurrences of the four inflections of the word. Wordforms originally unavailable in the source corpora (e.g., American and British English “inexistence” and “ice cream”) were intentionally assigned the frequency value “N.A.” on the basis of the following reasoning. First, a large portion of these words are multiword expressions, e.g., “city wall”, or hyphenated words, e.g., “t-shirt”, which are typically excluded from frequency lists (for instance, hyphenated words in SUBTLEX-UK were split into their individual components before counting word occurrences; van Heuven et al., 2014). Not only are they less relevant for most behavioral research, but they also pose significant challenges to the estimation of frequency values (see Gries, 2022, and O’Donnell, 2011 for overviews). Second,

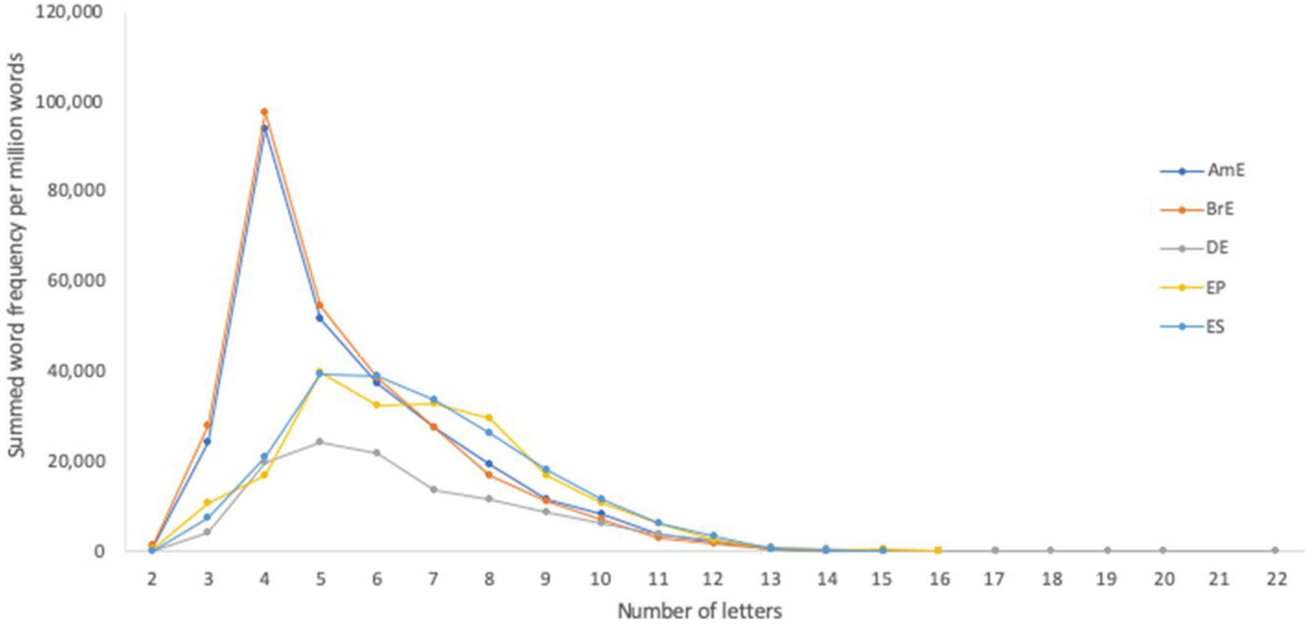


Fig. 1 PHOR-in-One summed corpus frequency distribution (per million words) by word length in each language. AmE = American English; BrE = British English; DE = German; EP = European Portuguese; ES = Spanish

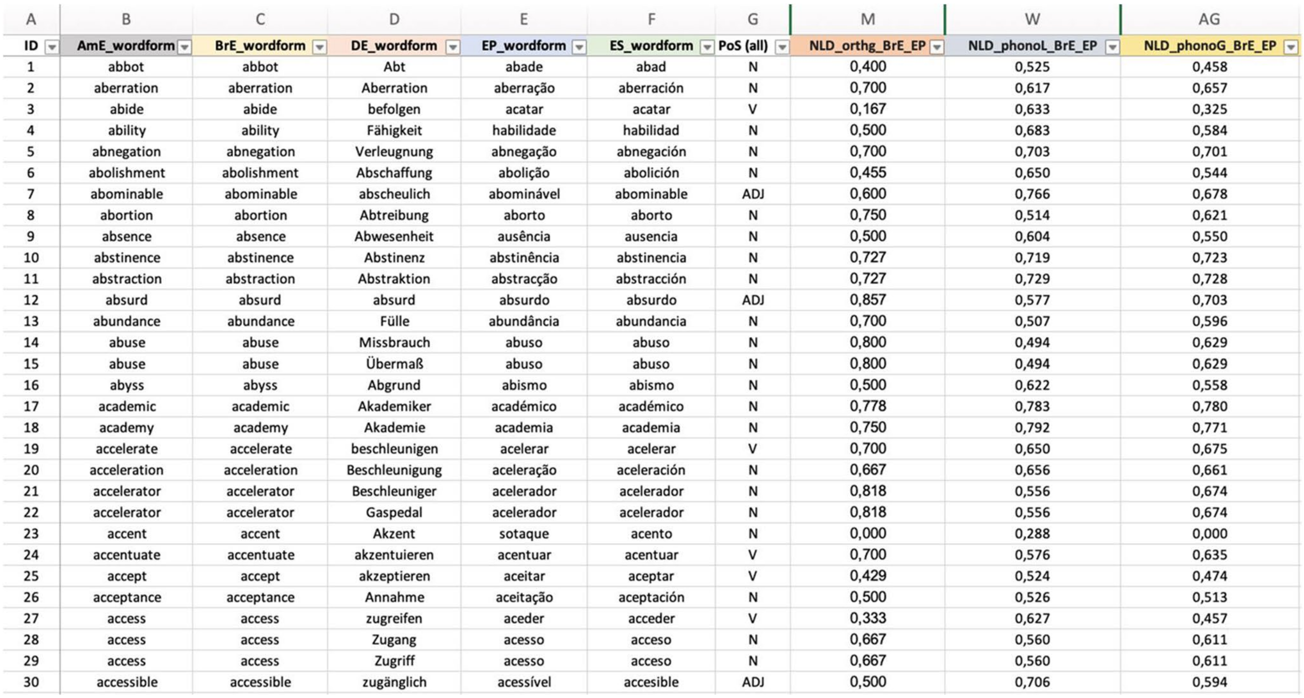
although a frequency of either zero or one is often indexed to non-occurring words (Brysbaert & Diependaele, 2013), these values may be an inaccurate approximation to the real frequency of the word. For instance, even though "*abril*" (European Portuguese for "April") has a fairly high (4.63) $\log_{10}$ corpus frequency in P-Pal (Soares et al., 2018a), surprisingly it does not occur in SUBTLEX-PT. In this case, a frequency of zero might mistakenly suggest that the word "*abril*" is highly infrequent and potentially unfamiliar to most speakers, when in fact it may simply reflect a more specific context of occurrence (P-Pal is essentially a newspaper corpus, which may increase the chances of certain word types appearing compared to subtitle corpora). Third, the simple words assigned a frequency of N.A. in PHOR-in-One make up for a very small portion of the lexica, ranging between 0.02% (Spanish) and 2.39% (American English) for corpus frequency, and 0.60% (British English and European Portuguese) and 3.20% (German) for subtitle frequency. Therefore, we opted for N.A. so that these few words would not interfere with any statistical computations based on lexical frequency in the database.

Figure 1 displays PoS-independent summed corpus frequency distribution as a function of word length for each language.

As illustrated, European Portuguese, German, and Spanish present a similar summed frequency distribution per word length, although the total summed frequencies in each language are much larger for European Portuguese and Spanish (40,858,792 and 63,338,916 tokens, respectively) than German (14,357,946 tokens). The German lexicon includes longer words (*max* = 22 letters) than the remaining languages (where maximum lengths range between 15 and 17 letters). The American and British English lexica have nearly overlapping frequency distributions, both peaking at four letters. All languages have a positively skewed distribution, suggesting that lexical frequency decreases as the number of letters in the word increase, as confirmed in previous works (e.g., Corral et al., 2015; Grzybek, 2007; Soares et al., 2014a, 2018a). For instance, Soares and collaborators (2018a) showed that approximately 55% of the European Portuguese lexical frequencies in P-PAL are accounted for by one, two, and three-letter words. However, in PHOR-in-One, 53% of European Portuguese frequencies are accounted for by five, six, and seven-letter words, presumably due to the fact that function words (e.g., pronouns and determiners), which typically capture a large percentage of the frequency of occurrence in a corpus (Brysbaert et al., 2012; Soares et al., 2014a, 2018a), were excluded here.

Structure of the database

PHOR-in-One is a read-only Excel file with one spreadsheet divided into three interconnected sections, each offering different types of information. The first section (columns A–G) displays five translation equivalents fully aligned in American English (AmE_wordform), British English (BrE_wordform), German (DE_wordform), European Portuguese (EP_wordform) and Spanish (ES_wordform), as well as their global PoS (PoS [all]), which is true for all languages at once. Each lexical entry is uniquely represented with an



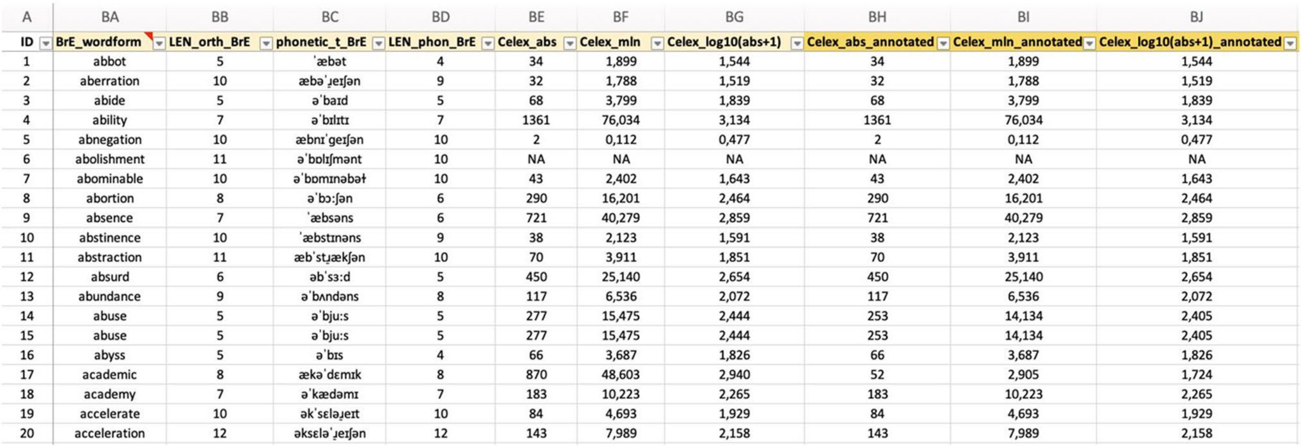
A	B	C	D	E	F	G	M	W	AG
ID	AmE_wordform	BrE_wordform	DE_wordform	EP_wordform	ES_wordform	PoS (all)	NLD_orthg_BrE_EP	NLD_phonoL_BrE_EP	NLD_phonoG_BrE_EP
1	abbot	abbot	Abt	abade	abad	N	0,400	0,525	0,458
2	aberration	aberration	Aberration	aberração	aberración	N	0,700	0,617	0,657
3	abide	abide	befolgen	acatar	acatar	V	0,167	0,633	0,325
4	ability	ability	Fähigkeit	habilidade	habilidad	N	0,500	0,683	0,584
5	abnegation	abnegation	Verleugnung	abnegação	abnegación	N	0,700	0,703	0,701
6	abolishment	abolishment	Abschaffung	abolição	abolición	N	0,455	0,650	0,544
7	abominable	abominable	abscheulich	abominável	abominable	ADJ	0,600	0,766	0,678
8	abortion	abortion	Abtreibung	aborto	aborto	N	0,750	0,514	0,621
9	absence	absence	Abwesenheit	ausência	ausencia	N	0,500	0,604	0,550
10	abstinence	abstinence	Abstinenz	abstinência	abstinencia	N	0,727	0,719	0,723
11	abstraction	abstraction	Abstraktion	abstracção	abstracción	N	0,727	0,729	0,728
12	absurd	absurd	absurd	absurdo	absurdo	ADJ	0,857	0,577	0,703
13	abundance	abundance	Fülle	abundância	abundancia	N	0,700	0,507	0,596
14	abuse	abuse	Missbrauch	abuso	abuso	N	0,800	0,494	0,629
15	abuse	abuse	Übermaß	abuso	abuso	N	0,800	0,494	0,629
16	abyss	abyss	Abgrund	abismo	abismo	N	0,500	0,622	0,558
17	academic	academic	Akademiker	académico	académico	N	0,778	0,783	0,780
18	academy	academy	Akademie	academia	academia	N	0,750	0,792	0,771
19	accelerate	accelerate	beschleunigen	acelerar	acelerar	V	0,700	0,650	0,675
20	acceleration	acceleration	Beschleunigung	aceleração	aceleración	N	0,667	0,656	0,661
21	accelerator	accelerator	Beschleuniger	acelerador	acelerador	N	0,818	0,556	0,674
22	accelerator	accelerator	Gaspedal	acelerador	acelerador	N	0,818	0,556	0,674
23	accent	accent	Akzent	sotaque	acento	N	0,000	0,288	0,000
24	accentuate	accentuate	akzentuieren	acentuar	acentuar	V	0,700	0,576	0,635
25	accept	accept	akzeptieren	aceitar	aceptar	V	0,429	0,524	0,474
26	acceptance	acceptance	Annahme	aceitação	aceptación	N	0,500	0,526	0,513
27	access	access	zugreifen	aceder	acceder	V	0,333	0,627	0,457
28	access	access	Zugang	acesso	acceso	N	0,667	0,560	0,611
29	access	access	Zugriff	acesso	acceso	N	0,667	0,560	0,611
30	accessible	accessible	zugänglich	acessível	accesible	ADJ	0,500	0,706	0,594

Fig. 2 PHOR-in-One lexical database (sections 1 and 2). AmE = American English; BrE = British English; DE = German; EP = European Portuguese; ES = Spanish. Depiction of the PHOR-in-One lexical database including five translation equivalents, (Columns B-F), their ID (Column A), global PoS (Column G), and orthographic (NLD_orthg_BrE_EP), phonological (NLD_phonoL_BrE_EP) and phonographic (NLD_phonoG_BrE_EP) NLD scores for British English-European Portuguese. Columns H-L, N-V and X-AF, containing form similarity estimates in other language combinations, are hidden

identity key (ID) specified in column A. The second section (columns H to AK) comprises orthographic (NLD_orthg), phonological (NLD_phonoL) and phonographic (NLD_phonoG) similarity scores for each language combination. Both sections are represented in Fig. 2, which displays lexical entries 1 to 30, their ID, five translation equivalents in American and British English, German, European Portuguese, and Spanish, PoS information and the orthographic, phonological, and phonographic similarity scores for British English-European Portuguese.

The third section offers a collection of linguistic information for each language individually. This includes number of letters (LEN_orth), phonetic transcription (phonetic_t) and number of phonemes (LEN_phon), as well as a set of PoS-independent and PoS-dependent absolute (abs), per-million-word (mln) and $\log_{10}$ of the absolute frequency + 1 (log10[abs+1]) corpus frequency indices (PoS-dependent frequency information is signaled with the use of the word "annotated" in the header, e.g., Celex_abs_annotated). The ensuing column (SpecPoS_annotated) presents a language-specific PoS tag, which will only be filled if the PoS-dependent corpus frequency estimates do not match the global PoS specified in column G. To illustrate, consider the American and British English wordform "accounting" in ID 42, and its translation equivalents "*Buchhaltung*", "*contabilidade*" and "*contabilidad*" in German, European Portuguese, and Spanish, respectively, all labeled as nouns (N) in column G (PoS [all]). Although it was possible to extract PoS-dependent frequencies for this wordform as a noun in the American English corpus (SpecPoS_annotated_ANC is blank), in British English the word only occurred as a verb. Therefore, the value (V) specified in column BK (SpecPoS_annotated_Celex) indicates that the annotated frequency values for British English (Celex_abs_annotated, Celex_mln_annotated and Celex_log10(abs+1)_annotated) are indexed to the PoS verb. Subsequently, SUBTLEX absolute (Subtlex_abs), per-million-word (Subtlex_mln), $\log_{10}$ absolute frequency + 1 (Subtlex_log10[abs+1]), and Zipf (Subtlex_zipf; as mentioned, Zipf subtitle frequencies are not available for German or Spanish) frequencies are displayed. The same properties are available across languages in the same order of presentation. Figure 3 depicts the linguistic and frequency information available for British English, as represented in PHOR-in-One.

For clarity, header labels are identical across languages, either preceded or followed by an abbreviation that specifies the corresponding source corpus (ANC, Celex, dlexDB, P-PAL or ESPAL) or language (AmE, BrE, DE, EP and ES for American English, British English, German, European Portuguese, and Spanish, respectively). A total of



A	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ
ID	BrE_wordform	LEN_orth_BrE	phonetic_t_BrE	LEN_phon_BrE	Celex_abs	Celex_mln	Celex_log10(abs+1)	Celex_abs_annotated	Celex_mln_annotated	Celex_log10(abs+1)_annotated
1	abbot	5	ˈæbət	4	34	1,899	1,544	34	1,899	1,544
2	aberration	10	æbəˈɹeɪʃən	9	32	1,788	1,519	32	1,788	1,519
3	abide	5	əˈbaɪd	5	68	3,799	1,839	68	3,799	1,839
4	ability	7	əˈbɪlɪtɪ	7	1361	76,034	3,134	1361	76,034	3,134
5	abnegation	10	æbnɪˈgeɪʃən	10	2	0,112	0,477	2	0,112	0,477
6	abolishment	11	əˈbɒlɪʃmənt	10	NA	NA	NA	NA	NA	NA
7	abominable	10	əˈbɒmɪnəbəɫ	10	43	2,402	1,643	43	2,402	1,643
8	abortion	8	əˈbɔ:ʃən	6	290	16,201	2,464	290	16,201	2,464
9	absence	7	ˈæbsəns	6	721	40,279	2,859	721	40,279	2,859
10	abstinence	10	ˈæbstɪnəns	9	38	2,123	1,591	38	2,123	1,591
11	abstraction	11	æbˈstɹækʃən	10	70	3,911	1,851	70	3,911	1,851
12	absurd	6	əbˈsɜ:d	5	450	25,140	2,654	450	25,140	2,654
13	abundance	9	əˈbʌndəns	8	117	6,536	2,072	117	6,536	2,072
14	abuse	5	əˈbju:s	5	277	15,475	2,444	253	14,134	2,405
15	abuse	5	əˈbju:s	5	277	15,475	2,444	253	14,134	2,405
16	abyss	5	əˈbɪs	4	66	3,687	1,826	66	3,687	1,826
17	academic	8	ækəˈdɛmɪk	8	870	48,603	2,940	52	2,905	1,724
18	academy	7	əˈkædəmɪ	7	183	10,223	2,265	183	10,223	2,265
19	accelerate	10	əkˈsɛləɹeɪt	10	84	4,693	1,929	84	4,693	1,929
20	acceleration	12	əksɛləˈɹeɪʃən	12	143	7,989	2,158	143	7,989	2,158

Fig. 3 PHOR-in-One lexical database (section 3). BrE = British English. Depiction of the third section of the PHOR-in-One lexical database including British English wordforms (BrE_wordform) for lexical entries 1-20 (Column A), and their length in number of letters (LEN_orth_BrE), phonetic transcription (phonetic_t_BrE), number of phonemes (LEN_phon_BrE), PoS-independent (Columns BE-BG) and PoS-dependent (Columns BH-BJ) corpus frequency estimates

111 columns are included. PHOR-in-One is available for download as an Excel file at https://www.psi.uminho.pt//pt/CIPsi/Laboratorios_Investigacao/Psicolinguistica/Documents/PHOR_in_One_LDB.zip and in the supplementary materials.

Results and discussion

Form similarity estimates

Phonological and orthographic similarity

Computation of the interlanguage phonological similarity estimates in PHOR-in-One required setting up an integrated multilingual phonetic alphabet, and the subsequent retrieval, standardization, and verification of phonetic transcriptions for all its lexical entries. Two principles underpinned the development of the phonetic alphabet: i) compliance with the notation of the IPA (International Phonetic Association, 1999), and ii) assignment of a single, unambiguous phonetic symbol to each sound across languages. Accordingly, some adjustments were made in our implementation of the phonological similarity algorithm, since the original version (Schepens, 2010) often employed identical characters to represent different sounds across languages (e.g., [r] represented the Spanish alveolar trill /r/, the English post-alveolar approximant /ɹ̠/, the German uvular trill /ʁ/, and the German post-vocalic /r/, [ɐ]; [R] represented both the Spanish simple alveolar tap/flap /ɾ/ and the British English post-alveolar approximant /ɹ̠/ at word endings; see Appendix Table 9 for details). To avoid the use of overlapping notations for different phonemes across languages, more specific phonetic segments were included in our version of the phonological similarity algorithm for each language. In addition, allophones, i.e., different realizations of the same phoneme, e.g., [l] and [ɫ], were also considered in our alphabet. For instance, the lateral approximant /l/ has two different realizations in British English: clear [l] at word-beginning and same-syllable pre-vocalic positions (e.g., "land" [lænd] and "plate" [pleɪt], respectively), and dark/velarized [ɫ] at word endings and before consonants (e.g., "full" [‘fʊɫ] and "belt" [‘bɛɫt], respectively). Conversely, in American English /l/ is always dark/velarized (e.g., "label" [‘ɫeɪbəɫ]). Although these phonetic specifications are typically not represented in dictionaries or psycholinguistic lexical databases, allophones of /l/, /t/, /d/, /b/, /g/, /r/, /n/, /m/ and /s/ were included to highlight different realizations of the same phoneme within and across languages, since recent evidence has suggested that bilingual individuals are sensitive to allophonic contrast (e.g., Burrows et al., 2019; Fabiano-Smith et al., 2015), and that allophones may in fact form the basis of pre-lexical processing during spoken-word recognition (e.g., Mitterer et al., 2018). Table 1 features the allophones incorporated into the PHOR-in-One multilingual alphabet.

Integration of the sounds of each language into a single multilingual inventory resulted in a phonetic alphabet that contains 37 pulmonic consonants and seven complex consonants, including four affricates (a cluster of two segments with the same place but different manners of articulation: [pf, ʦ, tʃ, dʒ]) and three co-articulated consonants (a cluster of segments with two simultaneous places of articulation: [kw, gw, ɣw]), distributed in the consonant space according to their articulatory features. The consonant matrix contains two overlapping tables (or dimensions), one for pulmonic consonants, and one for complex consonants, each with twelve columns and eight

Table 1 Allophone Inventory and contextual occurrence in PHOR-in-One for each language

IPA	Representation in PHOR-in-One	Language	Occurrence
/l/	clear [l]	DE	All positions, e.g., *Luft* [ˈlʊft]
		ES	All positions, e.g., *lucha* [ˈlutʃa]
		BrE	Word-beginning, e.g., land [ˈlænd]; same-syllable pre-vocalic positions, e.g., plate [ˈpleɪt]
	dark/velarized [ɫ]	EP	All positions, e.g., *língua* [ˈɫĩgwɐ]
		AmE	All positions, e.g., *land, plate, ful* [ˈɫænd, ˈpɫeɪt, ˈfʊɫ]
		BrE	Word endings, e.g., *full* [‘fʊɫ]; before consonants, e.g., *belt* [ˈbɛɫt]
/t/, /d/	flap [ɾ]	AmE	Between vowels, e.g., *battery* [ˈbæɾəɹ̠i]; after [ɹ̠], e.g., *article* [ˈɑɹ̠ɾɪkəɫ]
/r/	[ɹ̠]	AmE, BrE	All positions; used to differentiate from ES alveolar trill [r]
	[ɐ]	DE	Post-vocalic /r/, e.g., *Tier* [ˈti:ɐ]
/n/	[ɱ]	ES	Before [f], e.g., *confirmar* [koɱfiɾˈmaɾ]
/m/	[ɱ]	DE	Before [f], e.g., *Nymphe* [ˈnʏɱfə] and [v], e.g., *Invasion* [ɪɱva:ˈzjo:n]
		AmE, BrE	Before [f], e.g., *comfort* [ˈkʌɱfət, ˈkʌɱfəɹ̠t]
/d/	[ð]	EP	All positions except word beginnings and after nasal sounds, e.g., *acidente* [ɐsiˈðẽtɨ]
		ES	All positions except word beginnings, after [n] and [l], e.g., *accidente* [akθiˈðente]
/b/	[β]	EP	All positions except word beginnings and after nasal sounds, e.g., *abuso* [ɐˈβuzu]
		ES	All positions except word beginnings and after [m], e.g., *abuso* [aˈβus̺o]
/g/	[ɣ]	EP	All positions except word beginnings and after nasal sounds, e.g., *agudo* [ɐˈɣuðu]
		ES	All positions except word beginnings and after [n], e.g., *agudo* [ɐˈɣuðo]
/s/	apical [s̺]	ES	All positions; used to differentiate from EP laminal [s], e.g., *sótano* [ˈs̺otano]
	apical [z̺]	ES	Before voiced consonants, e.g., *eufemismo* [ewfeˈmiz̺mo]

AmE = American English; BrE = British English; DE = German; EP = European Portuguese; ES = Spanish

Table 2 Phonetic consonants adopted in PHOR-in-One and positioning in the IPA matrix reflecting their place and manner of articulation

	Labial		Coronal				Dorsal				Radical	Lar.
	Bilabial	Ld.	Dent	Alveolar	Palato-Alv.	Retro.	Alv.-Palat.	Palatal	Velar	Uvular	Pha+epi	glottal
Plosives	p b (pf)			t d (ʦ tʃ dʒ)					k g (kw gw)			
Nasals	m	ɱ		n				ɲ	ŋ			
Trills				r								
Taps or Flaps				ɾ								
Fricatives	β	f v (pf)	θ ð	s s̺ z z̺ (ʦ)	ʃ ʒ (tʃ dʒ)			ç ʝ	x ɣ (ɣw)	ʁ		h
L. Fricatives												
Approximants	w (kw gw ɣw)			ɹ̠				j				
L. Approximants				l ɫ			ʎ ʎ̟					

Note. *L*. Lateral; *Ld*. Labiodental; *Dent* Dental; *Retro*. Retroflex; *Palato-Alv*. Palato-Alveolar; *Alv*.-Palat. Alveolo-Palatal; *Pha* Pharyngeal; *Epi* Epiglottal; *Lar*. Laryngeal. Affricates and co-articulated consonants are represented between brackets in the lower section of each row and should be regarded as pertaining to a different consonant dimension (penalties are applied for substitutions involving such phonemes). Positioning of complex consonants is repeated in the matrix in order to reflect the place and/or manner of articulation of the two segments

rows that reflect their place (e.g., labial, coronal, dorsal) and manner (e.g., plosive, fricative, approximant) of articulation. These two overlapping dimensions are included in Table 2, which displays the positioning of each consonant in the IPA feature space (see also Appendix Table 10 for the full list of consonants adopted in PHOR-in-One).

The vowel inventory contains 39 (oral, nasal, short, long, and borrowed) vowels and 26 diphthongs, distributed in the IPA feature space according to their height (e.g., open, close) and backness (e.g., front, central). The vowel matrix contains four overlapping dimensions, for short vowels (oral and nasal, e.g., /i/ and /ũ/, respectively), long vowels (e.g.,

Table 3 Short (oral and nasal) vowels adopted in PHOR-in-One and positioning in the IPA matrix reflecting their height and backness

	Front	Near-Front	Central	Near-back	Back
Close	i ĩ		ɨ		u ũ
Close-Close-Mid		ɪ, ʏ		ʊ	
Close-Mid	e ẽ ø				o õ
Close-Mid-Open-Mid			ə		
Open-Mid	ɛ œ		ɜ		ʌ ɔ
Open-Mid-Open	æ		ɐ ɐ̃		
Open	a				ɑ ɒ

/iː/), borrowed vowels (e.g., /ɛ̃ː/), and diphthongs (e.g., /aɪ/), respectively. Table 3 displays (oral and nasal) short vowel positions in the IPA feature space. Long vowels, borrowed vowels and diphthongs are displayed in Table 4 (see also Appendix Table 11 for the full list of vowels and diphthongs adopted in PHOR-in-One, with examples).

Phonemic mergers were considered in our multilingual phonetic alphabet in order to standardize American English phonetic transcriptions. Due to the multiplicity of accents, phonological changes over time have resulted in large phonetic variability, complexifying the use of a representative phonetic notation (see Labov et al., 2008 and Hughes et al., 2012 for details). Hence, for parsimony, the following American English mergers were adopted in PHOR-in-One: i) hurry-furry: merge of /ʌr/ with /ɜr/; ii) horse-hoarse: merge of /ɔː/ and /oʊ/ (includes other word pairs such as war-wore and morning-mourning); iii) fern-fir-fur: merge of /ɛ, ɪ, ʊ/ into [ɜɹ] in coda positions; iv) cot-caught: merge of /ɔː/, /ɔ/ and /ɒ/; and e) intervocalic /ɒr/ merges with /ɑr/ and /ɔr/ (see Labov, 2006; Labov et al., 2008; and Wells, 1982 for an overview). The vowels adopted for American and British English are displayed in Appendix Table 12 with examples. Considering all these adjustments, overall, our multilingual phonetic alphabet includes 108 phonemes, 15 more than the original version (Schepens, 2010).

Upon the definition of the multilingual phonetic notation, the phonetic transcriptions retrieved from The Free Dictionary were standardized (e.g., [g] was converted to [g], and [ʧ] – one character – was converted to [tʃ] – two characters) and cross-checked using monolingual and bilingual dictionaries (European Portuguese: Casteleiro, 2001; German: Dudenredaktion, n.d., and Wiktionary; British English and American English: Oxford University Press, n.d.; Spanish: Real Academia Española, n.d. and Wiktionary), in addition to the following automatic phonetic converters online: Automatic Phonemic Transcriber, (Brondsted, n.d.), Res Publicae (Armario, 2008), Transcriptor Fonético (López, n.d.), Easy Pronunciation (Baytukalov, n.d.), tophonetics (Tophonetics, n.d.) and Text2Phonetics (Text2Phonetics, n.d.). If all transcriptions from these resources matched, they were automatically accepted as correct. When at least one of the transcriptions was different, they were verified by an expert in phonetics. European Portuguese transcriptions were retrieved from P-PAL (Soares et al., 2018a), or from The Free Dictionary and subsequently reviewed if unavailable.

Before computing the interlanguage phonological similarity measures, the phonetic transcriptions were automatically converted into an intermediate notation, so that complex characters like long vowels, e.g., [iː], and diphthongs, e.g., [əʊ], could be processed as units. For this purpose, an extended version of DISC and DISC++ (Schepens, 2010), a single-coded ASCII notation that visually resembles the IPA in a computer readable format (see the Celex

Table 4 Long vowels, borrowed vowels and diphthongs adopted in PHOR-in-One with positioning in the IPA matrix reflecting their height and backness

	Front	Near-Front	Central	Near-back	Back
Close	i: y: i:ɐ y:ɐ				u: u:ɐ
Close-Close-Mid		eɪ aɪ ɪɐ ɪ:ɐ ɪə ɔɪ ɔʏ ʏɐ		aʊ əʊ ʊɐ oʊ ʊə	
Close-Mid	e: ø: e:ɐ eɪ ø:ɐ				o: õ: o:ɐ oʊ
Close-Mid-Open-Mid			ɛə ɪə əʊ ʊə		
Open-Mid	ɛ: ɛ̃: ɛɐ ɛ:ɐ ɛə œɐ		ɜ:		ɔ: ɔʏ ɔɪ ɔɐ
Open-Mid-Open			i:ɐ y:ɐ e:ɐ ø:ɐ ɛɐ ɛ:ɐ œɐ aɐ a:ɐ ɪɐ ɪ:ɐ ʏɐ ʊɐ u:ɐ o:ɐ ɔɐ		
Open	a: ã aɐ a:ɐ aɪ aʊ				ɑ:

Note. Long and borrowed vowels are represented in the upper section of each row. Diphthongs (represented in the lower section of each row) are repeated in order to reflect the height and/or backness of both segments

English Linguistic Guide, 1995 for details) was created. Our extended version, DISC*, introduces new phonemes for American English and European Portuguese, and a set of allophones, as detailed in Tables 1 and 2 (see also Appendix Table 10 and 11 for the full IPA and DISC* alphabet with examples in each language).

To compute the interlanguage phonological similarity estimates, a modified version of the LD was implemented, in conjunction with an adaptation of the phoneme distance algorithm developed by Schepens (Schepens, 2010; see also Schepens et al., 2013). As mentioned, this algorithm is sensitive to phoneme qualities and modulates substitution costs according to the acoustic and articulatory features of the source and target phonemes. To estimate the cost of each substitution, it identifies the relative positions of the two phonemes in the consonant and vowel matrices of the IPA feature space (Tables 2, 3, and 4), and computes their Euclidean distance by applying Eq. (1), which determines the length of a line segment between them. The shorter the line, the smaller the distance.

$$\textit{Phoneme distance} = \frac{\sqrt{(\textit{column difference})^2 + (\textit{row difference})^2}}{\textit{Normalization constant}}. \tag{1}$$

By way of example, the distance between the velar plosive [k] and the glottal fricative [h] is 5. The minimum distance is 0, for phonemes with the same place and manner of articulation (e.g., [p] and [b]), whereas the maximum distance is 11.70, between a bilabial plosive ([p] or [b]) and the glottal fricative [h].

After computing the Euclidean distance between each source and target phoneme, penalties are added if at least one of the two is a long vowel, diphthong or borrowed vowel (for vowel substitutions), and a long affricate or co-articulated consonant (for consonant substitutions). The base penalty is set at 0.4 (e.g., substituting a short vowel with a diphthong, or vice-versa). Penalties are cumulative when neither the source or target phoneme is a short vowel or a pulmonic consonant. For instance, in the British English-German translation pair "analyse-analysieren" [ænəłaɪz-analy:zi:ʁən], the diphthong [aɪ] is replaced with the long vowel [y:], and hence two penalties are applied, producing a total penalty of 0.8. No penalties are applied between two short vowels, or two pulmonic consonants. The resulting phoneme distances, including penalties, are then divided by a normalization constant, so that individual phoneme substitution costs can be distributed between 0 and 2 (see Eq. 1). Upon testing six different values, a normalization constant of 5 was established for the implementation of the original algorithm, as it generated greater correlation coefficients with subjective ratings from previous studies (for further details see Schepens, 2010). Substitutions exceeding the maximum cost of 2 (e.g., the cost of substituting a bilabial plosive [p] or [b] with a glottal fricative [h] is 2.34) are automatically adjusted to 2. This cost redistribution between 0 and 2 is based on the premise that consonant/vowel substitutions are not allowed because they have distinct roles in word processing (e.g., Acha & Perea, 2010; Caramazza et al., 2000; Lee et al., 2002; Soares et al., 2020; Soares et al., 2014b). Therefore, in this algorithm, regarding substitutions as a deletion followed by an insertion, rather than a single operation, ensures that consonant/vowel substitutions instantly receive the maximum cost of 2 (Schepens, 2010).

Compared to the original version of the algorithm (Schepens, 2010), a major adjustment was carried out in our implementation, which impacted the computation of the phonological similarity scores. While insertions and deletions originally received a cost of one, here, those costs were adjusted to 2. The rationale behind this alteration was as follows. In the classical LD, every operation (substitution, insertion, and deletion) has an identical cost, i.e., 1. Given that phoneme substitutions here were set at a maximum cost of 2, insertions and deletions should also be adjusted to 2. In addition, due to the nature of the normalization formula in phonology, which multiplies the denominator by 2 as shown in Eq. (2), if insertion and deletion costs were set at 1 (as in the original proposal; Schepens, 2010), the resulting degree of phonological similarity would be overestimated, particularly for pairs involving multiple insertions or deletions.

$$\textit{phonological NLD} = \frac{\sum \textit{phoneme distances}}{\textit{Length of the longest string} * 2}. \tag{2}$$

To illustrate, consider the German-English pair "*Ende*-end" [ɛndə-ɛnd], with an orthographic NLD of .75. Like orthography, the source and target phonetic strings have four and three characters, respectively, the first three elements are identical, and one operation (insertion/deletion) is required to transform one string into the other. As such, in this pair, the orthographic and phonological NLDs should be identical, i.e., .75. While an insertion/deletion cost of 1 results in a phonological NLD of .88, a cost of 2 ensures identical NLD scores for orthography and phonology. Therefore, taken together, these arguments support our assumption that an insertion/deletion cost of 2 is more suitable for this algorithm than a cost of 1.

With these adjustments, our adaptation of the phonological Levenshtein distance algorithm, including the use of phoneme distances as substitution costs, can be defined as

$$lev_{a,b}\ (i,j) = \begin{cases} \max\ (i * k, j * k) & \text{if } \min\ (i,j) = 0, \\ \min \begin{cases} lev_{a,b}\ (i-1,j) + k_{a,b} \\ lev_{a,b}\ (i,j-1) + k_{a,b} \\ lev_{a,b}\ (i-1,j-1) + cost_{a,b_{(a_i \neq b_j)}} \end{cases} & \text{otherwise} \end{cases} \tag{3}$$

Table 5 Phonological Levenshtein distance matrix for the American/British English-European Portuguese pair "house-casa" using DISC*

	" "	h	6	s
" "	0	2	4	6
k	2	1	3	5
&	4	3	1.4	3.4
z	6	5	3.4	1.4
ɐ	8	7	5.4	3.4

The modified phonological Levenshtein distance implemented here considers the minimum number of substitutions, insertions, and deletions necessary to transform the phonetic transcription of the European Portuguese word "*casa*" [kazɐ] into the phonetic transcription of the English word "house" [haʊs], and vice-versa, using the Euclidean distance between the source and target phonemes as substitution costs. Substitution costs are distributed between 0 and 2, whereas insertion and deletion costs are set at 2. A single-coded DISC* notation is used for the computation, so that complex characters (e.g., diphthongs) can be interpreted as units

where $k = 2$[1]. For clarity, Table 5 displays the adapted LD matrix derived by our modified algorithm for the computation of the phonological distance between the American/British English-European Portuguese translation equivalents "house-*casa*" using the single-coded DISC* notation.

The value (3.4) displayed in the last entry corresponds to the final distance between the two strings, which is then normalized (see Eq. [2]). The resulting phonological NLD for this pair is .58.

As for orthography, the similarity scores were computed using the classical LD (Levenshtein, 1966). Table 6 displays the LD matrix for the computation of the orthographic distance between the American/British English-European Portuguese translation equivalents "house-*casa*" (the same as provided above for a direct comparison with phonology).

The value specified in the last entry of the matrix (4.00) corresponds to the final distance between the two strings, which is then normalized (Schepens et al., 2012). The resulting orthographic NLD for this pair is .20.

Table 7 displays the mean, median, maximum, and minimum orthographic and phonological NLD for each language pair in PHOR-in-One.

As illustrated, all pairs exhibit minimum and maximum orthographic NLDs of .00 and 1.00, respectively, indicating that orthographically entirely distinct and orthographically identical translation equivalents are included in the database. German-European Portuguese exhibits the smallest mean (.29) and median (.19) NLD, whereas European Portuguese-Spanish presents the largest (.77 and .86, respectively). As for phonological overlap, all language pairs have a minimum phonological NLD of .00, except for European Portuguese-Spanish (*min* = .12). British English-European Portuguese and British English-Spanish have a maximum phonological NLD of .98, indicating that there are no phonologically identical cognates for these language combinations in the database. The remaining language pairs have a maximum phonological NLD of 1.00. With the exception of European Portuguese-Spanish, mean phonological NLD scores are very close across language pairs, ranging between .49 for German-European Portuguese and .53 for British English-German. Conversely, mean orthographic NLD scores are more scattered, ranging between .29 for European Portuguese-German and German-Spanish, and .43 for British English-Spanish.

When orthographic and phonological similarity are taken together, European Portuguese-Spanish stands out for a number of reasons. First, 86% (N = 5,266) of European Portuguese-Spanish translation equivalents have an orthographic NLD score greater than or equal to .50, and the percentage is even higher for phonology (91%, N = 5,672), which signals the great formal proximity of the two languages. Second, not only are mean orthographic (.765) and phonological (.770) NLD scores considerably higher for this pair than for the other language combinations, they are also nearly identical. The fact that the mean difference between phonological and orthographic NLD scores for the remaining language pairs is substantially greater than zero (*Min* = .09 for British English-Spanish; *Max* = .22 for German-Spanish), suggests

[1] The adapted phonological Levenshtein distance algorithm applies Eq. (3), which stipulates two conditions to build a matrix m with i rows and j columns (see Table 5). The first condition (first line in Eq. [3]), sets the values for the first row (from 0 to 6 in the example from Table 5) and for the first column (from 0 to 8), and is fulfilled when the positions of either i or j are 0 (note that the first row and column correspond to position 0, e.g., [0,0] for row 0 and column 0, [0,1] for row 0 and column 1, [1,0] for row 1 and column zero, and so on). If the first condition is satisfied, each position is multiplied by constant k (with a value of 2), and the maximum value is selected. For instance, in the entry immediately below [h] in Table 5, at position (0,1), both 0 and 1 are multiplied by k, and 2 is selected to fill the entry. The second condition, specified in lines 2–4, performs three operations to set the values for the remaining positions in the matrix, where neither i or j are 0. The second line in Eq. (3) retrieves the value at $(i – 1, j)$ and computes the cost of performing an insertion by adding constant k. The third line retrieves the value at $(i, j – 1)$ and computes the cost of a deletion by adding k. Finally, the fourth line retrieves the value at $(i – 1, j – 1)$ and computes the cost of a substitution by adding the Euclidean distance + penalties (if applied). If the two phonemes are identical, the cost of a substitution is zero. Subsequently, the minimum value out of these computations is selected. Word transformation costs are computed incrementally. For instance, in Table 5, the value 3.4 in position (3,2) represents the transformation of [k&z] into [h6], and expresses the total cost of replacing [k] with [h] (substitution cost = 1.0), [&] with [6] (substitution cost = 0.4), and the cost of deleting [z] (deletion cost = 2.0). When comparing [k&z] with [h6s] at position (3,3), the cost of replacing [z] with [s] is zero, because the two phonemes have the same place and manner of articulation. As a result, transforming [k&z] into [h6s] has the same cost as that of transforming [k&] into [h6] (position [2,2] in Table 5), i.e., 1.4. The final phonological Levenshtein distance of the two translation equivalents corresponds to the value in the last entry of the matrix, i.e., 3.4, which reflects the cost of replacing [k] with [h], [&] with [6] and [z] with [s], and the cost of deleting [ɐ].

Table 6 Orthographic Levenshtein distance matrix for the English-European Portuguese pair "casa-house"

	" "	h	o	u	s	e
" "	0	1	2	3	4	5
c	1	1	2	3	4	5
a	2	2	2	3	4	5
s	3	3	3	3	3	4
a	4	4	4	4	4	4

The classical orthographic Levenshtein distance was implemented here, and considers the minimum number of substitutions, insertions, and deletions necessary to transform the European Portuguese word "*casa*" into the English word "house", and vice-versa. Substitution, insertion, and deletion costs are set at 1

that the phonological NLD scores are on average higher than the orthographic NLDs. A practical example is the British English-European Portuguese translation pair "veil-*véu*" [veɪɫ-vɛw], with an orthographic NLD of .25 and a phonological NLD of .77. This asymmetry is presumably due to the fact that, unlike orthography (where grapheme similarity does not play a role, and where substituting non-identical phonemes always has a cost of 1), many phoneme substitutions produce a cost which is smaller than 1, thus necessarily resulting in higher NLDs. The difference between individual orthographic and phonological NLD scores may be particularly pronounced for translation equivalents that involve fewer costly operations, such as insertions and/or deletions, or consonant-vowel substitutions. Hence, to compensate for this increment, Schepens (2010) and Schepens and collaborators (2013) proposed using a cognate inclusive threshold of .75 for phonology.

To explore the relationship between the two indices, Fig. 4 depicts a histogram for orthographic and phonological NLD scores in each language pair.

The distribution suggests that orthographic overlap is, in general, more positively skewed (see purple bars in Fig. 4), indicating that the amount of translation equivalents decreases as NLD intervals increase. Conversely, phonological NLD scores (pink bars) seem to fall into a bell-shaped distribution in most language pairs. European Portuguese-Spanish presents a different pattern, since orthography and phonology are both negatively skewed and have nearly overlapping distributions. A Kendall's Tau correlation analysis between orthographic and phonological NLDs was

Table 7 Mean, median, maximum, and minimum orthographic and phonological NLD scores in PHOR-in-One, and number of translation equivalents with orthographic NLD greater than or equal to .5 and phonological NLD greater than or equal to .75 with examples

Language Pair	Mean NLD (SD)	Median	Max NLD	Min NLD	Number of translation equivalents with NLD ≥ .5	Examples of translation equivalents with NLD ≥ .5
	Orthographic similarity					
BrE-DE	.375 (.316)	.250	1.000	.000	2,058	accent-*Akzent*
BrE-EP	.413 (.300)	.400	1.000	.000	2,828	alphabet-*alfabeto*
BrE-ES	.429 (.312)	.400	1.000	.000	2,877	accident-*accidente*
DE-EP	.286 (.269)	.194	1.000	.000	1,610	*Publikum-público*
DE-ES	.291 (.271)	.200	1.000	.000	1,626	*Kapazität-capacidade*
EP-ES	.765 (.264)	.857	1.000	.000	5,266	*aceitar-aceptar*
	Phonological similarity				Number of translation equivalents with NLD ≥ .75	Examples of translation equivalents with NLD ≥ .75
BrE-DE	.532 (.193)	.496	1.000	.000	1,034	[ˈæksənt-ak'ʦɛnt]
BrE-EP	.510 (.165)	.509	.980	.000	449	[ˈæɫfəbɛt- aɫfɐˈβɛtu]
BrE-ES	.520 (.178)	.513	.980	.000	706	[ˈæksɪdənt-akθiˈðente]
DE-EP	.492 (.155)	.467	1.000	.000	415	[ˈpu:blikʊm-ˈpuβɫiku]
DE-ES	.512 (.179)	.475	1.000	.000	844	[kapaʦiˈtɛ:t- kapaθiˈðað]
EP-ES	.770 (.160)	.797	1.000	.119	3,777	[ɐsɐjˈtaɾ-aθepˈtaɾ]

BrE = British English; DE = German; EP = European Portuguese; ES = Spanish. Language combinations involving American English were excluded, as they were similar to British English performed across all language pairs, as shown in Fig. 5 (for the sake of simplicity, language combinations involving American English were excluded, as they were very similar to British English).

As expected, all correlations were significant (all $p < .001$), given the large number of data points in the analysis. Considering orthographic and phonological similarity for the same language pairs (e.g., the correlation between orthographic and phonological NLD scores within British English-German, $t_b = .54$), only strong positive correlations were found (all t_b greater than or equal to .46), showing that phonological similarity increases as orthographic NLD scores increase. Strong positive correlations were also observed for different language pairs, namely between the orthographic NLD for British English-European Portuguese and the phonological NLD for British English-Spanish ($t_b = .50$), and between the orthographic NLD for German-European Portuguese and the phonological NLD for German-Spanish ($t_b = .47$). Interestingly, the correlation between the orthographic NLD for British English-European Portuguese and the phonological NLD for British English-Spanish ($t_b = .50$) is nearly identical to the correlation between the orthographic and phonological NLD within British English-European Portuguese ($t_b = .51$), potentially due to the close proximity of European Portuguese and Spanish orthography and phonology. The correlation between orthographic and phonological NLD scores within European Portuguese-Spanish is the strongest ($t_b = .59$) out of all comparisons, as expected.

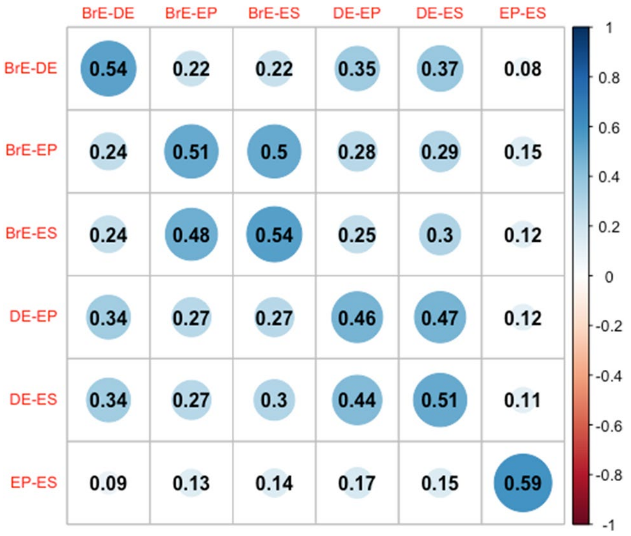
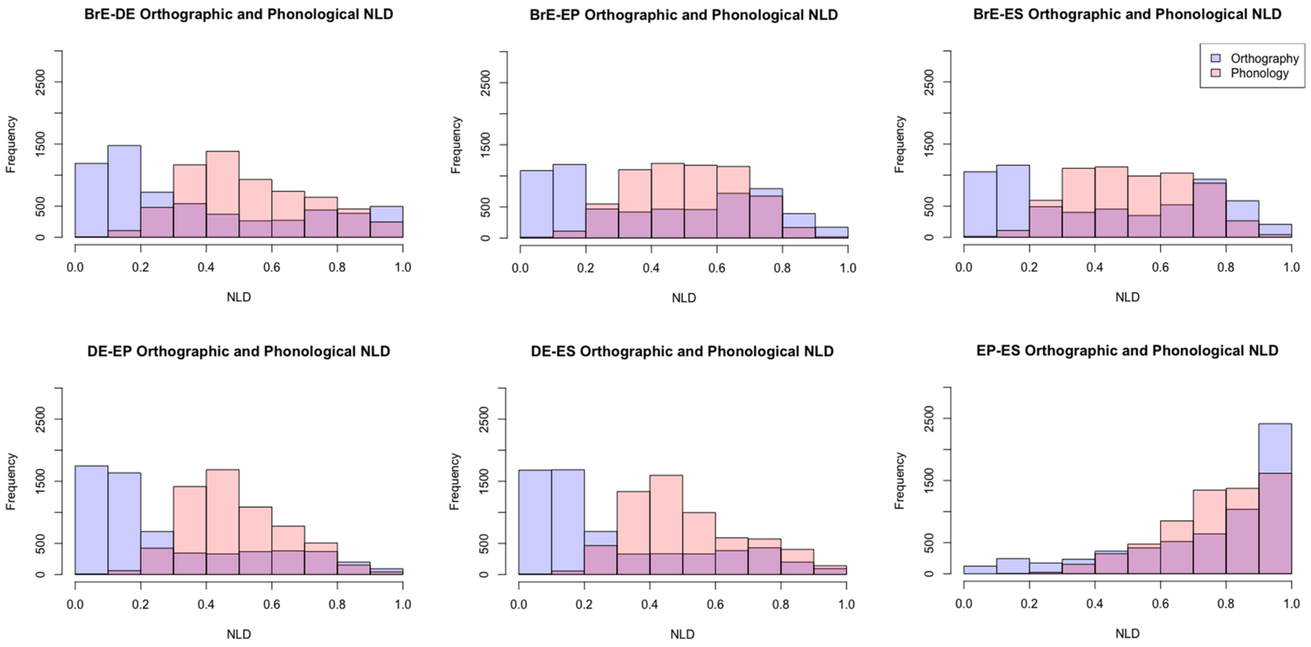


Fig. 5 Correlation matrix between the orthographic and phonological NLD scores across language combinations in PHOR-in-One. BrE = British English; DE = German; EP = European Portuguese; ES = Spanish. Blue circles in the correlogram denote positive correlations. Darker and larger circles denote stronger correlations. Correlation coefficients range from .08 (for the correlation between the orthographic NLD scores in British English-German and the phonological NLD scores in European Portuguese-Spanish, and vice-versa) and .59 (for the correlation between the orthographic and phonological NLD scores in European Portuguese-Spanish). All $p < .001$. Correlations involving American English were excluded, as they were very similar to British English


BrE-DE Orthographic and Phonological NLD

BrE-EP Orthographic and Phonological NLD

BrE-ES Orthographic and Phonological NLD

DE-EP Orthographic and Phonological NLD

DE-ES Orthographic and Phonological NLD

EP-ES Orthographic and Phonological NLD

Fig. 4 Histogram of the distribution of orthographic and phonological NLD scores in each language pair. BrE = British English; DE = German; EP = European Portuguese; ES = Spanish. Language combinations involving American English were excluded as they were very similar to British English

The distribution of orthographic and phonological NLD scores for each language pair presented in this section shows that PHOR-in-One allows for the selection of stimuli with distinct formal features, including, i) orthographically and phonologically identical cognates, e.g., British English-German “lift-*Lift*” [lɪft-lɪft] (orthographic NLD = 1.00; phonological NLD = 1.00); ii) orthographically, but not phonologically related cognates, e.g., British English-German “psychologist-*Psychologe*” [saɪkɒlədʒɪst-psy:çolo:gə] (orthographic NLD = .75; phonological NLD = .41); iii) phonologically, but not orthographically related cognates, e.g., English-German “ice-*Eis*” [aɪs-aɪs] (orthographic NLD = .00, phonological NLD = 1.00); and iv) orthographically and phonologically distinct translation pairs, or noncognates, e.g., American English-Spanish “bear-*oso*” [bɛəɹ̠-os̪o] (orthographic NLD = .00; phonological NLD = 0.36). This variability is important for research, and is an adequate representation of how the words are distributed in a language (Siew & Vitevitch, 2019).

Phonographic similarity

A new objective index of interlanguage phonographic overlap is also introduced in PHOR-in-One. The definition of a phonographic NLD should satisfy the following guiding principles: i) express the degree of overall form similarity of two translation equivalents by intersecting their orthographic and phonological overlap; ii) ensure an intuitive categorization of translation pairs as low or high-similarity; iii) be distributed on a continuum between .00 and 1.00 for comparability with other form similarity estimates; iv) approximate the mean of the orthographic and phonological NLD scores. To apply these principles the geometric mean of the individual orthographic and phonological NLD scores was computed. We opted for the geometric rather than the arithmetic mean because it tends to dampen the effects of high values (Habib, 2012), thus levelling the differences reported above between the two measures. However, due to the nature of the geometric mean, which uses the product of n values, the resulting phonographic NLD is zero when at least one of the values is zero. For instance, in the British English-German translation equivalents ice-*Eis* [aɪs-aɪs], with an orthographic NLD of 0.00 and a phonological NLD of 1.00, the phonographic NLD using the geometric mean is zero. This is a serious limitation, in that it does not capture any information about non-zero values (de la Cruz & Kreft, 2019) and violates principle iv). To address this issue, an extension of the geometric mean was implemented here, which can handle zero values efficiently, and which has been used before in other scientific areas (e.g., Alexander et al., 2005; Williams, 1937). In this extension, 1 is added to individual NLD scores, before estimating the product of the orthographic and phonological NLD, and subsequently subtracted from the result, as expressed in (4),

Table 8 Number of translation equivalents in four phonographic NLD intervals for each language combination in PHOR-in-One with examples

Languages	NLD = .000	.000 < NLD < .500	.500 ≤ NLD < 1.000	NLD = 1.000
	Two languages			
BrE-DE	1 poor-*arm*	3952 acorn-*Eichel*	2190 activity-*Aktivität*	17 film-*Film*
BrE-EP	3 aunt-*tia*	3317 accent-*sotaque*	2840 ability-*habilidade*	0 NA
BrE-ES	3 nail-*uña*	3283 advantage-*ventaja*	2874 absence-*ausencia*	0 NA
DE-EP	0 NA	4459 *Schauspieler-actor*	1701 *Adoption-adopção*	0 NA
DE-ES	1 *Topf-olla*	4434 *Tier-animal*	1718 *Agonie-agonía*	7 *Mango-mango*
EP-ES	0 NA	783 *adicionar-añadir*	5318 *acesso-acceso*	59 *flor-flor*
	Three languages			
BrE-DE-EP	0 NA	2532 addiction-*Sucht-vício*	1451 active-*aktiv-activo*	0 (NA)
BrE-DE-ES	0 NA	2495 wall-*Wand-pared*	1464 insulin-*Insulin-insulina*	0 (NA)
BrE-EP-ES	0 NA	616 barn-*celeiro-granero*	2610 lemon-*limão-limón*	0 (NA)
DE-EP-ES	0 NA	692 *Hochzeit-casamento-boda*	1558 *Vulkan-vulcão-volcán*	0 (NA)
	Four languages			
BrE-DE-EP-ES	0 NA	485 window-*Fenster-janela-ventana*	1350 vein-*Vene-veia-vena*	0 (NA)

BrE = British English; DE = German; EP = European Portuguese; ES = Spanish; NA = Not available. Language combinations involving American English were excluded, as they were similar to British English

$$G(X) = \left(\prod_{i=1}^{n} \left(x_i + 1 \right) \right)^{\frac{1}{n}} - 1, \quad (4)$$

where $x \geq 0$.

Table 8 details the total number of translation equivalents in PHOR-in-One with a phonographic NLD of .000 (noncognates with no orthographically and phonologically overlapping features), 1.000 (orthographically and phonologically identical words), and distributed between .001 and .499 and .500 and .999 with examples for two, three and four language combinations.

The numbers show that hardly any translation equivalents with a phonographic NLD of .00 exist in PHOR-in-One. Only one pair of such translations is included for British English-German (“poor-*arm”* [ˈpʊə - ˈa:ɐm]) and German-Spanish (“*Topf-olla*” [ˈtɔpf - ˈoʝa]), and three for British English-European Portuguese (e.g., “aunt-*tia*” [ˈɑ:nt - ˈtiɐ]) and British English-Spanish (e.g., “nail-*uña*” [ˈneɪɫ - ˈuɲa]). The remaining language combinations do not contain phonographically non-overlapping translation equivalents. Additionally, only British English-German (e.g., “film-*Film*” [ˈfɪɫm - ˈfɪlm]), German-Spanish (e.g., “*Mango-mango*” [ˈmaŋgo - ˈmaŋgo]) and European Portuguese-Spanish (e.g., “*flor-flor*” [ˈfɫoɾ - ˈfloɾ]) contain phonographically identical cognates (17, 7, and 59 translation equivalents, respectively). The European Portuguese-Spanish pair shares a larger number of phonographically similar (i.e., .50 ≤ NLD <1.00; e.g., “*acesso-acceso*”) and identical (NLD = 1.00) words than any other language combination (*N* = 5377 combined). When three language combinations are considered at once, a large number of phonographically distinct (i.e., .00 ≤ NLD < .50; *min* = 616 words in British English-European Portuguese-Spanish, e.g. “barn [ˈbɑ:n] - *celeiro* [sɨˈɫɐjɾu] - *granero* [gɾaˈneɾo]”; *max* = 2532 words in British English-German-Spanish, e.g., “wall [wɔ:ɫ] - *Wand* [vant] - *pared* [paɾeð]”) and phonographically similar (i.e., .50 ≤ NLD < 1.00; *min* = 1451 words in British English-German-Spanish, e.g., “insulin [ˈɪnsjʊlɪn] - *Insulin* [ɪnzuˈli:n] - *insulina* [ins̠uˈlina]”; *max* = 2610 words in British English-European Portuguese-Spanish, e.g., “lemon [ˈlɛmən] - *limão* [ɫiˈmɐ̃w] - *limón* [liˈmon]”) translation equivalents are part of the lexicon. A total of 1350 phonographically similar (e.g., “vein [ˈveɪn] - *Vene* [ˈve:nə] - *veia* [ˈvɐjɐ] - *vena* [ˈbena]”) and 485 phonographically distinct (e.g., “window [ˈwɪndəʊ] - Fenster [ˈfɛnstɐ] - janela [ʒɐˈnɛɫɐ] - Ventana [benˈtana]”) translation equivalents across the four languages at once are also included in PHOR-in-One.

To further assess how the phonographic NLD is distributed across languages, Fig. 6 depicts a histogram of the number of translation equivalents in each interval for six language pairs.

Compared to the histograms in Fig. 4, where for most language pairs orthographic NLD peaks between .00 and .20 and phonological NLD between .30 and .50 (except for

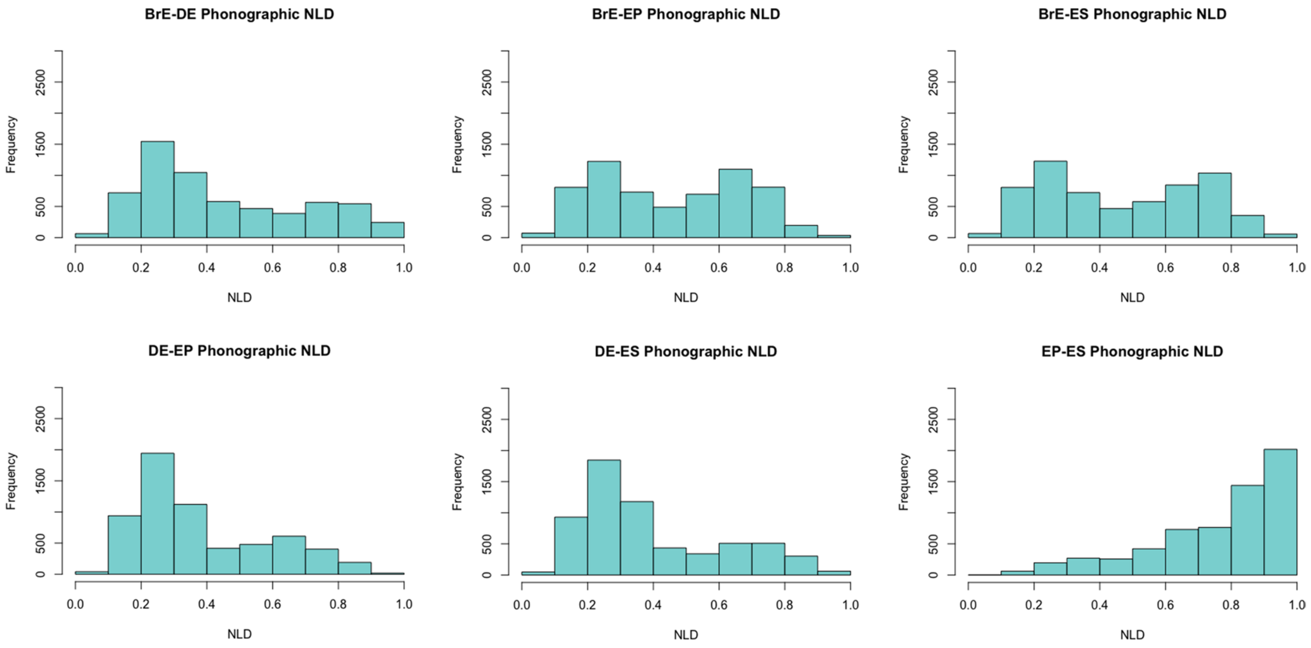


Fig. 6 Histogram of the distribution of phonographic NLD scores in each language pair. BrE = British English; DE = German; EP = European Portuguese; ES = Spanish. Language combinations involving American English were excluded as they were very similar to British English

British English-European Portuguese and British English-Spanish, both peaking between .30 and .70, and European Portuguese-Spanish, which presents a different distribution), phonographic NLD generally peaks between .20 and .30, suggesting that most translation equivalents in these language pairs share few orthographic and phonological features at once. British English-European Portuguese and British English-Spanish seem to be approximately binormally distributed, peaking between .20 and .30 and also between .60 and .70 (for British English-European Portuguese) or .70 and .80 (for British English-Spanish). European Portuguese-Spanish, however, remains negatively skewed, with a greater concentration of translation equivalents displaying a phonographic NLD between .90 and 1.00. Overall, compared to the orthographic and phonological NLD distributions in Figure 4, the histograms in Figure 6 reflect the fact that the phonographic NLD is a more intermediate index of similarity. To illustrate, the British English-European Portuguese translation equivalents "veil-*véu*" [veɪɫ-vɛw], which, as mentioned, have an orthographic and phonological NLD of .25 and .77, respectively, bear a phonographic NLD of .49.

The differences between the orthographic and phonological NLD scores signal the importance of controlling for both orthographic and phonological overlap in the selection of cognates and noncognates. In addition, the phonographic NLD may help calibrate these differences between the two measures, and potentially account for more variance in bilingual and multilingual performances, particularly for translation pairs with contrasting degrees of orthographic and phonological similarity. This is therefore a more straightforward method for assessing the degree of overall form similarity of two words, with particular relevance for researchers interested in selecting translation equivalents with high (O+P+) or low (O-P-) degrees of orthographic and phonological similarity.

Conclusions

In this paper we introduced PHOR-in-One, an extensive multilingual lexical database containing 6160 translation equivalents fully aligned in American and British English, German, European Portuguese and Spanish, as well as their linguistic, morphosyntactic and frequency characterization. To address a long-needed research requirement, PHOR-in-One offers three indices of interlanguage form similarity, including the classical orthographic NLD, an adapted phonological NLD, which considers the degree of proximity of the source and target phonemes in the IPA feature space as substitution costs, and an estimate of phonographic NLD, a simplified alternative to control for and/or manipulate orthographic and phonological similarity at once. Combined, these indices allow for the selection of comprehensive stimulus sets, including orthographically and phonologically distinct translation equivalents (or noncognates), orthographically but not phonologically related translation equivalents, phonologically but not orthographically related translation equivalents, and orthographically and phonologically identical cognates. PHOR-in-One will promote the adoption of comparable estimates of interlanguage form similarity across studies, increase speed and reliability in the process of stimulus selection, and contribute to expand the predictions of bilingual computational models on phonological processing and representation. Future studies should test how well the new indices of form similarity proposed here can capture subjects' performances with different tasks, word types and populations.

Appendix

Table 9 DISC++ Consonants, Corresponding Representation in DISC* and IPA and context of use in PHOR-in-One

Language	DISC++	DISC*	IPA	Use in PHOR-in-One
ES	/r/	/r̲/	/r/	Alveolar trill; word beginnings (“*rico*”), medial positions (“*terraza*”)
BrE	/r/	/r/	/ɹ̠/	Post-alveolar approximant; word beginnings (“rat”), syllable onset (“grass”)
	/r/			Excluded in coda positions due to assimilation (“arm”)
	/R/			Excluded in word-ending positions due to assimilation (“affair”)
DE	/r/	/Ʀ/	/ʁ/	Uvular trill/guttural R; word beginnings (“*Roman*”), syllable onset positions (“*Brot*”)
	/r/	/ɐ/	/ɐ/	Post-vocalic [r] (“*Tier*”)
ES	/R/	/ɾ/	/ɾ/	Simple alveolar tap/flap (“*harina*”)
BrE	/l/	/l/	/l/	Lateral approximant; word beginnings (“land”), same-syllable pre-vocalic positions (“plate”)
		/ɫ/	/ɫ/	Lateral approximant; dark/velarized [ɫ]: word endings (“full”) and before consonants (“belt”)
DE	/l/	/l/	/l/	Lateral approximant. Clear [l]: all positions (“*Luft*”*,* “*begleiten*”*,* “*Engel*”)
ES	/l/	/l/	/l/	Lateral approximant. Clear [l]; all positions (“*libro*”*,* “*ángel*”*,* “*amplitude*”)
ES	/L/	/Ɉ/	/ʝ/	Voiced palatal fricative; all positions (“*tobillo*”*,* “*playa*”); *yeísmo* is adopted
ES	/d/	/d/	/d/	Voiced alveolar plosive; word beginnings (“*desayuno*”), after ⟨n⟩ and ⟨l⟩ (“*espalda*”)
		/D/	/ð/	Allophone of /d/; remaining positions (“*absurdo*”)
ES	/b/	/b/	/b/	Voiced bilabial plosive; word beginnings (“*brazo*”) and after ⟨m⟩ (“*cambio*”)
		/B/	/β/	Allophone of /b/; remaining positions (“*abril*”*,* “*active*”)
ES	/g/	/g/	/g/	Voiced velar plosive; word beginnings (“*goma*”) and after ⟨n⟩ (“*sangre*”)
		/G/	/γ/	Allophone of /g/; remaining positions (“*hormiga*”)
ES	/s/	/ṣ/	/s̺/	Voiceless alveolar fricative, apical (“*sótano*”); all positions except before voiced consonants
ES		/ʐ/	/z̺/	Allophone of /s/, apical; before voiced consonants (“*abismo*”*,* “*disgusto*”)
DE	/x/	/x/	/x/	Voiceless velar fricative; after back vowels (“*Sucht*”*,* “*Koch*”) and after /a, a:/ (“*lachen*”*,* “*Nachteil*”)
	/x/	/ç/	/ç/	Voiceless palatal fricative; after front vowels (“*Bericht*”*,* “*Knöchel*”) and consonants (“*Mädchen*”)
AmE		/r/	/ɹ̠/	Post-alveolar approximant; all positions (“rat”, “trust”, “alter”)
AmE		/ɾ/	/ɾ/	Simple alveolar tap/flap; allophone of /t/ and /d/ between a stressed and unstressed vowel (“butter”, “body”), between two unstressed vowels (“ability”), and after /ɹ̠/ (“artifact”)
AmE		/ɫ/	/ɫ/	Dark/velarized [ɫ]; all positions (“land”, “plate”, “full”)
EP		/Ʀ/	/ʁ/	Uvular trill/guttural R; word beginnings (“*romance*”) and medial positions with ⟨rr⟩ (“*carro*”)
EP		/ɾ/	/ɾ/	Simple alveolar tap/flap (“*farinha*”)
EP		/l/	/l/	Dark/velarized [ɫ]; all positions (“*livro*”*,* “*adulto*”*,* “*abril*”)
EP		/ɏ/	/ʎ̟/	Alveolo-palatal lateral approximant; all positions (“*abelha*”)
EP		/d/	/d/	Voiced alveolar plosive; word beginnings (“*desafio*”) and after nasal sounds (“*amêndoa*”)
EP		/D/	/ð/	Allophone of /d/; remaining positions (“*ácido*”)
EP		/b/	/b/	Voiced bilabial plosive; word beginnings (“*braço*”) and after nasal sounds (“*combate*”)
EP		/β/	/β/	Allophone of /b/; remaining positions (“*abril*”)
EP		/g/	/g/	Voiced velar plosive; word beginnings (“*garrafa*”) and after nasal sounds (“*sangue*”)
EP		/G/	/γ/	Allophone of /g/; remaining positions (“*Agosto*”)
ES, DE, BrE, AmE		/ɱ/	/ɱ/	Voiced labiodental nasal; allophone of /m/. ES: before ⟨f⟩ (“*énfasis*”); DE: before ⟨f⟩ and ⟨v⟩ in simple words (“*Konflikt*”*,* “*Konvention*”); BrE and AmE: before ⟨ph⟩, ⟨mf⟩ and ⟨mv⟩ (“emphasis”, “comfort”; no occurrences in PHOR-in-One)

AmE = American English; BrE = British English; DE = German; EP = European Portuguese; ES = Spanish. Identical characters were originally adopted in DISC++ to represent different sounds within and across languages. The character /r/ was used to represent Spanish /r/, English /ɹ̠/, and German /ʁ/ and /ɐ/. The character /x/ was used to represent the German voiceless velar fricative /x/, and the voiceless palatal fricative /ç/. In the DISC* extension, each sound has a single, unambiguous phonetic representation (/r̲/, /r/,/Ʀ/, /ɐ/, /x/ and /ç/, respectively). Furthermore, allophones /ɫ/, /γ/, /s̺/, /z̺/, and /ɱ/ were added in DISC*. Spanish apical /s/ [s̺] and /z/ [z̺] were adopted to differentiate from laminal realizations in American and British English, German, and European Portuguese (note that these specifications will not affect the computation of phonological similarity, since both realizations have the same place and manner of articulation; the same is true for /l/ and /ɫ/). Blank cells refer to phonemes that were not included, either in the original study (Schepens, 2010) or in PHOR-in-One. Examples in each language are provided between brackets

Table 10 Phonetic Alphabet Adopted in PHOR-in-One and Corresponding DISC, DISC++ and DISC* Phonetic Codes (Consonants)

					Examples from PHOR-in-One				
Category	IPA	DISC	DISC++	DISC*	EP	ES	DE	BrE	AmE
Plosive	p	p	p	p	pacifismo	pacifismo	Pazifismus	pacifism	pacifism
	b	b	b	b	botânico	botánico	botanisch	botanical	botanical
	t	t	t	t	tabaco	tabaco	Tabak	tobacco	tobacco
	d	d	d	d	discussão	discusión	Diskussion	discussion	discussion
	k	k	k	k	caiaque	kayak	Kajak	kayak	kayak
	ɡ	g	g	g	glossário	glosario	Glossar	glossary	glossary
Nasal	m	m	m	m	magnetismo	magnetismo	Magnetismus	magnetism	magnetism
	n	n	n	n	nação	nación	Nation	nation	nation
	ɲ		µ	µ	cunhado	cuñado			
	ŋ	N	N	N		ancla	Anker	anchor	anchor
	ɱ			ɱ		confirmar	Invasion	lymph, comfort	lymph, comfort
Fricative	f	f	f	f	fanático	fanático	fanatisch	fanatic	fanatic
	v	v	v	v	vulcânico		vulkanisch	volcanic	volcanic
	β			β	baba	observar			
	θ	T	T	T		crecimiento		growth	growth
	ð	D	D	D	abade	madre		mother	mother
	s	s	s	s	servidor		Server	server	server
	s̺			ṣ		sostener			
	z	z	z	z	zumbido		Summen	buzz	buzz
	z̺			ʐ		desviar			
	ʃ	S	S	S	champanhe		Champagner	champagne	champagne
	ʒ	Z	Z	Z	jornalismo		Journalismus	collision	collision
	ç	x	x	ç			Mechaniker		
	x	x	x	x		vigilancia	Überwachung		
	ʁ	r		Ʀ	romance		Roman		
	γ		G	G	tigre	tigre			
	ʝ			Ɉ		ayuda, calle			
	h	h	h	h			Hotel	hotel	hotel
Approx.	ɹ̠	r	r	r				rat	rat, arm
	w	w	w	w	acção	actual		warning	warning
	j	j	j	j	aceitar	abierto	Aktion	abuse	abuse
(lateral)	ɫ			ɫ	legião			abominable	legion, abominable
	l	l	l	l		legión	Legion	legion	
	ʎ̟			ɏ	alho				
Tap / Flap	ɾ		R	ɾ	desviar	desviar			
Trill	r		r	r̠		rico			
Affricate	pf	+	+	+			Pferd		
	ʦ	=	=	=			Zusatz		
	tʃ	J	£	J		cochero	Kutscher	coachman	coachman
	dʒ	_	¥	_			Dschungel	angel	angel
Co-articulated consonants	kw		¤	¤	quantia	acuerdo		acquire	acquire
	ɡw			È̈	guarnição	guapo		iguana	iguana
	γw			ű	água	desguace			

Approx. = Approximant; AmE = American English; BrE = British English; DE = German; EP = European Portuguese; ES = Spanish

DISC and DISC++ characters were generally maintained in DISC*. New phonemes were added using characters that visually resemble the IPA symbol as much as possible. Affricates [tʃ] and [dʒ] have different representations in DISC and DISC++ ([J] and [£], respectively), and hence the original DISC representation was maintained. The European Portuguese phonetic alphabet contains 19 consonants, three co-articulated consonants [kw, γw, gw], three allophones [β, ð, γ] and two semivowels [j, w]. The Spanish phonetic alphabet contains 26 consonants including four allophones [β, ð, γ, ɱ], three co-articulated consonants [kw, γw, gw], one affricate [tʃ], and two semivowels [j, w]. The narrower characters [s̺] and [z̺] are used to distinguish the Spanish apical /s/ from the American and British English, German and European Portuguese laminal /s/. German contains 22 consonants including one allophone [ɱ] and four affricates [pf, ts, tʃ, dʒ]. British English and American English include 24 consonants, two co-articulated consonants [kw, gw] and two affricates [tʃ, dʒ]. Note that both varieties include a dark/velarized [ɫ], but only British English includes clear [l]. American English includes the alveolar tap/flap [ɾ] as an allophone of /t, d/

Table 11 Phonetic alphabet adopted in PHOR-in-One and corresponding DISC, DISC++ and DISC* phonetic codes (vowels and diphthongs)

					Examples from PHOR-in-One				
Categ.	IPA	DISC	DISC++	DISC*	EP	ES	DE	BrE	AmE
Long	i:	i	I	i			Bibel	ease	
	y:	y	y	y			Bühne		
	e:	e	e	e			Athlet		
	ø:	|	|	|			Böse		
	a:	a	a	a			Kajak		
	o:	o	o	o			Legion		
	u:	u	u	u			akut	room, lose, group	
	ɑ:	#	#	#				palm, pass	
	ɔ:	$	$	$				law, altar	
	ɜ:	3	3	3				absurd, bird	
	ɛ:	)	)	)			Kapazität		
Short	a	&	&	&	álcool	alcohol	Alkohol		
	ɑ	A	A	A					palm, wash, law, altar
	ɒ	Q	Q	Q				wash	
	æ	{	{	{				cat, marry	cat, marry, pass
	ɐ			ɐ	aluno		Schüler		
	ə	@	@	@			Bibel	winner, error	winner, error
	ɜ			Ǝ					absurd, bird
	ɛ	E	E	E	cepticismo		Skepsis	bed, meadow	bed, meadow
	e			ê	etiqueta	etiqueta	Etikett		
	i			í	centímetro	centímetro	Zentimeter	ambience	ease, ambience, geography
	ɪ	ɪ	ɪ	ɪ			Diskussion	geography	
	ʏ	ʏ	ʏ	ʏ			Unglück		
	ɨ			^	desviar				
	ø			ø			Möblierung		
	œ	/	/	/			göttlich		
	ɔ	O	O	O	cosmos		Kosmos		horse
	o			ô	historiador	historiador	Historiker		
	ʌ	V	V	V				run, enough, brother	run, enough, brother
	ʊ	U	U	U			Kunst	put, foot	put, foot
	u			ú	unir	unir	Humanismus		room, lose, group
Nasal	ã			ã			Orange		
	ɛ̃:			Ĕ			Bulletin		
	õ			õ	concluir				
	ũ			ũ	renunciar				
	ẽ			ẽ	crente				
	ĩ			ĩ	ninfa				
	ɐ̃			å	âncora				
	õ:			ỡ			Saison		

Table 11 (continued)

					Examples from PHOR-in-One				
Categ.	IPA	DISC	DISC++	DISC*	EP	ES	DE	BrE	AmE
Dthg.	eɪ	1	1	1				bay, weight	bay, weight
	aɪ	2	2	2			Offenheit	light, apply, rice	light, apply, rice
	ɔɪ	4	4	4				noise, boy	noise, boy
	aʊ	6	6	6			Haus	house, allow	house, allow
	ɔʏ	X	X	X			Freund		
	a:ɐ			ạ			Arm		
	aɐ			ả			hart		
	ɛ:ɐ			Ē			Bär		
	ɛɐ			e̢			Termin		
	e:ɐ			ḛ			Verkehr		
	i:ɐ			ï			Tier, Vampir		
	y:ɐ			€			Tür, Gebühr		
	u:ɐ			%			Agentur, Uhr		
	ɪ:ɐ			]			Bildschirm		
	ɪɐ			;			Viertel		
	ʏɐ			ẏ			Kürze		
	ʊɐ			ṷ			Kurs		
	ɔɐ			ɷ			Norden		
	œɐ			ϱ			Körper		
	ø:ɐ			ǿ			Friseur		
	əʊ	5	5	5				bingo, know, toecap	
	o:ɐ			ǫ			Tor		
	oʊ			,					bingo, know, toecap
	ɛə	8		8				hair, premiere	
	ɪə	7		7				appear, seriousness	appear, seriousness
	ʊə	9		9				contour, poor, jury	contour, poor, jury

Note. Categ. = Category; AmE = American English; BrE = British English; DE = German; EP = European Portuguese; ES = Spanish; Dthg. = Diphthong. Original DISC and DISC++ characters were generally maintained in DISC*. New phonemes were added using characters that visually resemble the IPA symbol as much as possible. Multiple examples are provided to illustrate different grapheme-to-phoneme conversions, e.g., /aɪ/ in "light, apply, rice". The European Portuguese phonetic alphabet contains 14 short vowels, five of which are nasal sounds. Diphthongs are formed from vowel/semivowel combinations (e.g., [aj]), and are thus not represented in the table. Spanish contains five short oral vowels. Similar to European Portuguese, diphthongs are formed from vowel/semivowel combinations (e.g., [aj]), and are thus also not represented in the table. German contains 25 vowels, including long and short segments and three nasal vowels [õ:, ɛ̃:, ã]. In addition to diphthongs [aɪ, aʊ, ɔʏ], the uvular [ʁ] is vocalized as [ɐ] in post-vocalic positions (e.g., "*Arm*"), before consonants (e.g., "*Viertel*"), word-final positions (e.g., "*Bär*"), and ⟨-er⟩ word endings (e.g., "*Tier*"), resulting in a total of 19 diphthongs. British English contains 13 vowels, five of which are long, and eight diphthongs. American English contains 10 short vowels and seven diphthongs. Note that the British English segment [əʊ] is represented using [oʊ] in American English

Table 12 Vowels and diphthongs adopted for BrE and AmE in PHOR-in-One

BrE	AmE	Examples	BrE phonetic transcription	AmE phonetic transcription
ɑː	ɑ	father, palm, arm	[fɑ:ðə, pɑ:m, ɑ:m]	[fɑðəɹ, pɑm, ɑɹm]
ɑː	æ	slanderer, pass, path	[slɑ:ndəɹə, pɑ:s, pɑ:θ]	[słændəɹəɹ, pæs, pæθ]
ɒ	ɑ	dialogue, boss, contrast, qualify	[daɪəlɒg, bɒs, kɒntɹɑ:st, kwɒlɪfaɪ]	[daɪəłɑg, bɑs, kɑntɹæst, kwɑłɪfaɪ]
ɔː	ɑ	law, altar, applause	[lɔ:, ɔ:łtə, əplɔ:z]	[łɑ, ɑłtəɹ, əpłɑz]
ɒɹ	ɔɹ	horror, orange, quarrelsome, warrior	[hɒɹə, ɒɹɪndʒ, kwɒɹəłsəm, wɒɹɪə]	[hɔɹəɹ, ɔɹəndʒ, kwɔɹəłsəm, wɑɹjəɹ]
ɔː	ɔɹ	hoarse, horse, north	[hɔ:s, hɔ:s, nɔ:θ]	[hɔɹs, hɔɹs, nɔɹθ]
ɛə	ɛəɹ	hair	[hɛə]	[hɛəɹ]
iː	i	ease, see, siege, ceiling	[i:z, si:, si:dʒ, si:lɪŋ]	[iz, si, sidʒ, siłɪŋ]
ɪə	ɪəɹ	year	[jɪə]	[jɪəɹ]
ɪ	i	mania, geography	[meɪnɪə, dʒɪɒɡɹəfɪ]	[meɪniə, dʒiɑɡɹəfi]
əʊ	oʊ	go, hope, know, poem	[gəʊ, həʊp, nəʊ, pəʊɪm]	[goʊ, hoʊp, noʊ, poʊɪm]
ʊə	ʊɹ, ʊəɹ	tourism, endure	[tʊəɹɪzəm, ɪndjʊə]	[tʊɹɪzəm, ɛndjʊəɹ]
uː	u	loser, soon, acoustics	[lu:zə, su:n, əku:stɪks]	[łuzəɹ, sun, əkustɪks]
ʌɹ	ɜɹ	furrow, courage, thorough	[fʌɹəʊ, kʌɹɪdʒ, θʌɹə]	[fɜɹoʊ, kɜɹɪdʒ, θɜɹoʊ]
ɜː	ɜɹ	furniture, bird, word, absurd, advert	[fɜ:nɪtʃə, bɜ:d, wɜ:d, əbsɜ:d, ædvɜ:t]	[fɜɹnɪtʃəɹ, bɜɹd, wɜɹd, əbsɜɹd, ædvɜɹt]
ə	əɹ	winner, error, secular, confirmation	[wɪnə, ɛɹə, sɛkjʊlə, kɒnfəmeɪʃən]	[wɪnəɹ, ɛɹəɹ, sɛkjʊłəɹ, kɑnfəɹmeɪʃən]
ju:	ju	abuse, accumulation	[əbju:s, əkju:mjʊleɪʃən]	[əbjus, əkjumjəłeɪʃən]
ɛ	ɛ	bed, egg, meadow, error	[bɛd, ɛg, mɛdəʊ, ɛɹə]	[bɛd, ɛg, mɛdoʊ, ɛɹəɹ]
i	i	ambience, aquarium, calorie	[ˈæmbiəns, əkwɛəɹiəm, kæləɹi]	[æmbiəns, əkwɛəɹiəm, kæłəɹi]
ɪ	ɪ	bit, will, roses, affected, mirror	[bɪt, wɪł, əfɛktɪd, mɪɹə]	[bɪt, wɪł, əfɛktɪd, mɪɹəɹ]
ɔɪ	ɔɪ	boy, noise	[bɔɪ, nɔɪz]	[bɔɪ, nɔɪz]
ʊ	ʊ	put, foot, wolf	[pʊt, fʊt, wʊłf]	[pʊt, fʊt, wʊłf]
ʌ	ʌ	run, enough, mother, bathtub	[ɹʌn, ɪnʌf, mʌðə, bɑ:θtʌb]	[ɹʌn, ənʌf, mʌðəɹ, bæθtʌb]
ə	ə	roundabout, appointment, balance	[ɹaʊndəbaʊt, əpɔɪntmənt, bæləns]	[ɹaʊndəbaʊt, əpɔɪntmənt, bæłəns]
aɪ	aɪ	price, pie, daylight, kayak	[pɹaɪs, paɪ, deɪlaɪt, kaɪæk]	[pɹaɪs, paɪ, deɪłaɪt, kaɪæk]
aʊ	aʊ	house, tower	[haʊs, taʊə]	[haʊs, taʊəɹ]
æ	æ	cat, marry, parallel	[kæt, mæɹɪ, pæɹəlɛł]	[kæt, mæɹi, pæɹəłɛł]
eɪ	eɪ	daylight, pain, weight	[deɪlaɪt, peɪn, weɪt]	[deɪłaɪt, peɪn, weɪt]
ɪə	ɪə	adherence, editorial, idealism	[ədhɪəɹəns, ɛdɪtɔ:ɹɪəł, aɪdɪəlɪzəm]	[ədhɪəɹəns, ɛdɪtɔɹɪəł, aɪdiəłɪzəm]
jʊ	jʊ	articulate, valuation	[ɑ:ˈtɪkjʊleɪt, væljʊeɪʃən]	[ɑɹtɪkjʊłeɪt, væłjueɪʃən]

Note. AmE = American English; BrE = British English. Long vowels are only included in the British English alphabet. American English same-syllable post-vocalic /r/ is rhotic, whereas British English /r/ is not (Hosseinzadeh, Kambuziya & Shariati, 2015)

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-022-01985-3.

Author note We have no known conflict of interest to disclose. This study was conducted at the Psychology Research Centre (PSI/01662), School of Psychology, University of Minho, and was supported by the Foundation for Science and Technology (FCT) through the Portuguese State Budget (Ref.: UIDB/PSI/01662/2020) and grant PD/BD/52396-2013. We would like to thank Job Schepens for kindly sharing the original algorithm to compute the Euclidean phoneme distance, Tânia Senra for her valuable help with the review of the lexical entries, and Anabela Rato for guidance in the early stages of this study. We would also like to thank Miguel Marques for providing computer programming assistance.

References

Acha, J., & Perea, M. (2010). On the role of consonants and vowels in visual-word processing: Evidence with a letter search paradigm. *Language and Cognitive Processes, 25*(3), 423–438. https://doi.org/10.1080/01690960903411666

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

Alexander, N. D., Solomon, A. W., Holland, M. J., Bailey, R. L., West, S. K., Shao, J. F., Mabey, D. C., & Foster, A. (2005). An index of community ocular Chlamydia trachomatis load for control of trachoma. *Transactions of the Royal Society of Tropical Medicine* *and Hygiene, 99*(3), 175–177. https://doi.org/10.1016/j.trstmh.2004.05.003

Ando, E., Jared, D., Nakayama, M., & Hino, Y. (2014). Cross-script phonological priming with Japanese Kanji primes and English targets. *Journal of Cognitive Psychology, 26*(8), 853–870. https://doi.org/10.1080/20445911.2014.971026

Arana, S., Oliveira, H., Fernandes, A. I., Soares, A. P., & Comesaña, M. (2022). Does the cognate effect depend on the proportion of identical cognates? A study with Portuguese-English bilinguals. *Bilingualism, Language and Cognition, 25*(4), 660–678. https://doi.org/10.1017/S1366728922000062

Armario, J. (2008). *Res Publicae*. Retrieved July 1, 2020, from http://www.respublicae.net/lengua/silabas/

Arnon, I., & Christiansen, M. H. (2017). The Role of Multiword Building Blocks in Explaining L1–L2 Differences. *Topics in Cognitive Science, 9*(2), 621–636. https://doi.org/10.1111/tops.12271

Baayen, R. H., Piepenbrock, R., & L. Gulikers. (1995). The Celex Lexical Database (Release2) {CD-ROM}. *Linguistic Data Consortium*, University of Pennsylvania, Philadelphia, U.S.A.

Baytukalov, T. (n.d.). *Easy Pronunciation*. Retrieved July 1, 2020, from https://easypronunciation.com/en/english-phonetic-transcription-converter/

Blumenfeld, H. K., & Marian, V. (2005). Covert bilingual language activation through cognate word processing: An eye-tracking study. *Proceedings of the XXVII Annual Meeting of the Cognitive Science Society (Stresa), 27*, 286–291.

Brenders, P., van Hell, J. G., & Dijkstra, T. (2011). Word recognition in child second language learners: Evidence from cognates and false friends. *Journal of Experimental Child Psychology, 109*, 383–396. https://doi.org/10.1016/j.jecp.2011.03.012

Broersma, M., Carter, D., & Acheson, D. J. (2016). Cognate costs in bilingual speech production: Evidence from language switching. *Frontiers in Psychology, 7*, 1461. https://doi.org/10.3389/fpsyg.2016.01461

Brondsted, T. (n.d.). *Automatic Phonemic Transcriber*. Retrieved July 20, 2020, from http://tom.brondsted.dk/text2phoneme/?viewevalval&l=German

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

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

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

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

Burrows, L., Jarmulowicz, L., & Oller, D. K. (2019). Allophony in English Language Learners: The Case of Tap in English and Spanish. *Language, Speech and Hearing Services in Schools, 50*(1), 138–149. https://doi.org/10.1044/2018_LSHSS-17-0081

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*, 1209–1229. https://doi.org/10.1007/s11145-018-9835-8

Caramazza, A., Chialant, D., Capasso, R., & Miceli, G. (2000). Separable processing of consonants and vowels. *Nature, 403*(6768), 428–430. https://doi.org/10.1038/35000206

Casteleiro, J. M. (dir.). (2001). *Dicionário da Língua Portuguesa Contemporânea*. [Dictionary of the contemporary Portuguese Language]. : Academia das Ciências de Lisboa/Editorial Verbo.

CELEX English Linguistic Guide (1995). Retrieved June 16, 2021, from https://catalog.ldc.upenn.edu/docs/LDC96L14/eug_let.pdf

Christoffels, I. K., de Groot, A. M. B., & Kroll, J. F. (2006). Memory and Language Skills in Simultaneous Interpreters: The Role of Expertise and Language Proficiency. *Journal of Memory and Language, 54*(3), 324–345. https://doi.org/10.1016/j.jml.2005.12.004

Clifton, C. (2015). The Roles of Phonology in Silent Reading: A Selective Review. In: Frazier, L., Gibson, E. (eds) *Explicit and Implicit Prosody in Sentence Processing. Studies in Theoretical Psycholinguistics, 46*. Springer, Cham. https://doi.org/10.1007/978-3-319-12961-7_9

Comesaña, M., Ferré, P., Romero, J., Guasch, M., Soares, A. P., & García-Chico, T. (2015). Facilitative effect of cognate words vanishes when reducing the orthographic overlap: The role of stimuli list composition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 614–635. https://doi.org/10.1037/xlm0000065

Comesaña, M., Sánchez-Casas, R., Soares, A. P., Pinheiro, A. P., Rauber, A., Frade, S., & Fraga, I. (2012). The interplay of phonology and orthography in visual cognate word recognition: An ERP study. *Neuroscience Letters, 529*(1), 75–79. https://doi.org/10.1016/j.neulet.2012.09.010

Corral, Á., Boleda, G., & Ferrer-i-Cancho, R. (2015). Zipf's Law for Word Frequencies: Word Forms versus Lemmas in Long Texts. *PLoS ONE, 10*(7), e0129031. https://doi.org/10.1371/journal.pone.0129031

Costa, A., Caramazza, A., & Sebastián-Gallés, N. (2000). The Cognate Facilitation Effect: Implications for Models of Lexical Access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1283–1296. https://doi.org/10.1037//0278-7393.26.5.1283

Costa, A., Santesteban, M., & Caño, A. (2005). On the facilitatory effects of cognate words in bilingual speech production. *Brain and Language, 94*(1), 94–103. https://doi.org/10.1016/j.bandl.2004.12.002

Cuetos, F., Glez-Nosti, M., Barbon, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica, 32*, 133–143.

Cunningham, T. H., & Graham, C. R. (2000). Increasing native English vocabulary recognition through Spanish immersion: Cognate transfer from foreign to first language. *Journal of Educational Psychology, 92*(1), 37–49. https://doi.org/10.1037/0022-0663.92.1.37

de Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning, 50*(1), 1–56. https://doi.org/10.1111/0023-8333.00110

de Groot, A. M. B., & Nas, G. L. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language, 30*(1), 90–123. https://doi.org/10.1016/0749-596X(91)90012-9

de la Cruz, R., & Kreft, J-U., (2019). Geometric mean extension for data sets with zero. Available online https://arxiv.org/abs/1806.06403. https://doi.org/10.48550/arXiv.1806.06403

Dijkstra, T., Grainger, J., & van Heuven, W. J. B. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language, 41*(4), 496–518. https://doi.org/10.1006/jmla.1999.2654

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language, 62*(3), 284–301. https://doi.org/10.1016/j.jml.2009.12.003

Dijkstra, T., & Van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition, 5*(3), 175–197. https://doi.org/10.1017/S1366728902003012

Dijkstra, T., Wahl, A., Buytenhuijs, F., van Halem, N., Al-Jibouri, Z., de Korte, M., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition, 22*(4), 657–679. https://doi.org/10.1017/S1366728918000287

Dimitropoulou, M., Duñabeitia, J. A., & Carreiras, M. (2011). Phonology by itself: Masked phonological priming effects with and without orthographic overlap. *Journal of Cognitive Psychology, 23*(2), 185–203. https://doi.org/10.1080/20445911.2011.477811

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods, 45*(4), 1246–1258. https://doi.org/10.3758/s13428-013-0326-1

Dudenredaktion (n.d.). *Duden Online*. Retrieved July 1, 2020, from https://www.duden.de/woerterbuch/

Duyck, W. (2005). Translation and associative priming with cross-lingual pseudohomophones: evidence for nonselective phonological activation in bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1340–1359. https://doi.org/10.1037/0278-7393.31.6.1340

Duyck, W., Desmet, T., Verbeke, L., & Brysbaert, M. (2004). WordGen: a tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments & Computers, 36*(3), 488–499. https://doi.org/10.3758/BF03195595

Fabiano-Smith, L., Oglivie, T., Maiefski, O., & Schertz, J. (2015). Acquisition of the stop-spirant alternation in bilingual Mexican Spanish–English speaking children: Theoretical and clinical implications. *Clinical Linguistics & Phonetics, 29*(1), 1–26. https://doi.org/10.3109/02699206.2014.947540

Ferré, P., Sánchez-Casas, R., Comesaña, M., & Demestre, J. (2017). Masked translation priming with cognates and noncognates: Is there an effect of words' concreteness? *Bilingualism: Language and Cognition, 20*(4), 770–782. https://doi.org/10.1017/S1366728916000262

Gollan, T. H., Forster, K., & Frost, R. (1997). Translation priming with different scripts: masked priming with cognates and noncognates in Hebrew-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(5), 1122–1139. https://doi.org/10.1037//0278-7393.23.5.1122

Gries, S. T. (2022). Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach. *Lexis, 19*. https://doi.org/10.4000/lexis.6231

Grzybek, P. (2007). History and Methodology of Word Length Studies. In P. Grzybek (eds) *Contributions to the Science of Text and Language. Text, Speech and Language Technology, 31*, 15–90. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-4068-9_2

Guasch, M., Boada, R., Ferré, P., & Sánchez-Casas, R. (2013). NIM: A Web-based Swiss Army knife to select stimuli for psycholinguistic studies. *Behavior Research Methods, 45*(3), 765–771. https://doi.org/10.3758/s13428-012-0296-8

Habib, E. A. E. (2012). Geometric mean for negative and zero values. *International Journal of Research and Reviews in Applied Sciences, 11*(3), 419–432.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB – A lexical database for the psychological and linguistic research. *Psychologische Rundschau, 62*(1), 10–20. https://doi.org/10.1026/0033-3042/a000029

Holmes, J., & Ramos, R. G. (1993). False friends and reckless guessers: Observing cognate recognition strategies. In T. Huckin, M. Haynes, & J. Coady (Eds.), *Second Language Reading and Vocabulary Learning* (pp. 86–108). Ablex.

Hoshino, N., & Kroll, J. F. (2008). Cognate effects in picture naming: Does cross-language activation survive a change of script? *Cognition, 106*(1), 501–511. https://doi.org/10.1016/j.cognition.2007.02.001

Hosseinzadeh, N. M., Kambuziya, A. K. Z., & Shariati, M. (2015). British and American Phonetic Varieties. *Journal of Language Teaching and Research, 6*(3), 647–655. https://doi.org/10.17507/jltr.0603.23

Hughes, A. D., Trudgill, P., & Watt, D. (2012). *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles*. (5th ed.). Routledge. https://doi.org/10.4324/9780203784440

Ide, N. (2009). The American National Corpus: Then, Now, and Tomorrow. In Michael Haugh et al. (Eds.), *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, 108–113.

Iniesta, A., Rossi, E., Bajo, M. T., & Paolieri, D. (2021). The Influence of Cross-Linguistic Similarity and Language Background on Writing to Dictation. *Frontiers in Psychology, 12*, 679956. https://doi.org/10.3389/fpsyg.2021.679956

International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge, U.K: Cambridge University Press.

Labov, W. (2006). *The Social Stratification of English in New York City* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511618208

Labov, W., Ash, S., & Boberg, C. (2008). *The Atlas of North American English*. Mouton De Gruyter. https://doi.org/10.1515/9783110167467

Lee, H., Rayner, K., & Pollatsek, A. (2002). The processing of consonants and vowels in reading: Evidence from the fast priming paradigm. *Psychonomic Bulletin & Review, 9*, 766–772. https://doi.org/10.3758/BF03196333

Lemhöfer, K., & Dijkstra, T. (2004). Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition, 32*(4), 533–550. https://doi.org/10.3758/BF03195845

Lemhöfer, K., Dijkstra, T., & Michel, M. C. (2004). Three languages, one ECHO: Cognate effects in trilingual word recognition. *Language and Cognitive Processes, 19*(5), 585–611. https://doi.org/10.1080/01690960444000007

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics – Doklady, 10*, 707–710.

Lieber, R. (2010). *Introducing morphology*. Cambridge University Press.

López, X. (n.d.). *Transcriptor Fonético*. Retrieved July 20, 2020, from https://xavierlopez.dev/transcriptorfonetico/

Lotto, L., & de Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning, 48*(1), 31–69. https://doi.org/10.1111/1467-9922.00032

Marian, V. (2017). Orthographic and Phonological Neighborhood Databases across Multiple Languages. *Writ Lang Lit., 20*(1), 7–27.

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE, 7*(8), e43230. https://doi.org/10.1371/journal.pone.0043230

Mitterer, H., Reinisch, E., & McQueen, J. M. (2018). Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language, 98*, 77–92. https://doi.org/10.1016/j.jml.2017.09.005

Moon, R. (2015). Multi-word items. In John R. Taylor (Ed.), *Oxford Handbook of the Word* (pp. 120). Oxford University Press.

Nespor, M., Shukla, M., & Mehler, J. (2011). Stress-Timed vs. Syllable-Timed Languages. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology*. (Wiley-Blackwell), 1147–1159.

Nuerk, H. C., Rey, A., Graf, R., & Jacobs, A. M. (2000). Phonographic sublexical units in visual word recognition. *Current Psychology Letters, 2*, 25–36. https://doi.org/10.4000/cpl.241

O'Donnell, M. B. (2011). The adjusted frequency list: A method to produce cluster-sensitive frequency lists. *ICAME Journal, 35*, 135–169.

Oxford University Press. (n.d.). *Oxford Advanced Learner's Dictionary*. Retrieved July 01, 2020, from https://www.oxfordlearnersdictionaries.com/

Peereman, R., & Content, A. (1997). Orthographic and phonological neighborhoods in naming: Not all neighbors are equally influential in orthographic space. *Journal of Memory and Language, 37*(3), 382–410. https://doi.org/10.1006/jmla.1997.2516

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition, 40*(3), 551–577. https://doi.org/10.1017/S0272263117000407

Poort, E. D., & Rodd, J. M. (2019). A Database of Dutch–English Cognates, Interlingual Homographs and Translation Equivalents. *Journal of Cognition, 2*(1), 15. https://doi.org/10.5334/joc.67

Post da Silveira, A., & van Leussen, J. W. (2015). Generating a bilingual lexical corpus using interlanguage normalized Levenshtein distances. In *Proceeding of the 18th International Conference of Phonetic Sciences (XVII ICPhS)*, Glasgow, UK

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. https://doi.org/10.1017/S1366728915000206

Real Academia Española. (n.d.). *Diccionario de la lengua española*, 23.ª ed., [versión 23.5 en línea]. Retrieved July 01, 2020, from https://dle.rae.es/

Schepens, J. (2010). *Cross-Language Distributions of High Frequency and Phonetically Similar Cognates.* [Unpublished Master's Thesis]. Radboud University, Nijmegen, The Netherlands

Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition, 15*, 157–166. https://doi.org/10.1017/S1366728910000623

Schepens, J., Dijkstra, T., Grootjen, F., & van Heuven, W. J. B. (2013). Cross-Language Distributions of High Frequency and Phonetically Similar Cognates. *PLoS ONE, 8*(5), e63006. https://doi.org/10.1371/journal.pone.0063006

Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, and Psychophysics, 74*(1), 5–35. https://doi.org/10.3758/s13414-011-0219-2

Schwartz, A. I., Kroll, J. F., & Diaz, M. (2007). Reading words in Spanish and English: Mapping orthography to phonology in two languages. *Language and Cognitive Processes, 22*(1), 106–129. https://doi.org/10.1080/01690960500463920

Siew, C. S. Q., & Vitevitch, M. S. (2019). The phonographic language network: Using network science to investigate the phonological and orthographic similarity structure of language. *Journal of Experimental Psychology: General, 148*(3), 475–500. https://doi.org/10.1037/xge0000575

Soares, A. P., Iriarte, A., Almeida, J. J., Simões, A., Costa, A., Machado, J., et al. (2018a). Procura-PALavras (P-PAL): A web-based interface for a new European Portuguese lexical database. *Behavior Research Methods, 50*(4), 1461–1481. https://doi.org/10.3758/s13428-018-1058-z

Soares, A. P., Lages, A., Silva, A., Comesaña, M., Sousa, I., Pinheiro, A. P., & Perea, M. (2019a). Psycholinguistic variables in visual-word recognition and pronunciation of European Portuguese words: A megastudy approach. *Language, Cognition and Neuroscience, 4*(6), 689–719. https://doi.org/10.1080/23273798.2019.1578395

Soares, A. P., Machado, J., Costa, A., Iriarte, A., Simões, A., Almeida, J. J., Comesaña, M., & Perea, M. (2014a). On the advantages of frequency measures extracted from subtitles: The case of Portuguese. *Quarterly Journal of Experimental Psychology, 68*(4), 1–41. https://doi.org/10.1080/17470218.2014.964271

Soares, A. P., Oliveira, H. M., Comesaña, M., & Costa, A. S. (2018b). Lexico-syntactic interactions in the resolution of relative clause ambiguities in a second language (L2): The role cognate status and L2 proficiency. *Psicológica, 39*, 164–197. https://doi.org/10.2478/psicolj-2018-0008

Soares, A. P., Oliveira, H. M., Ferreira, M., Comesaña, M., Macedo, A. F., Ferré, P., …. & Fraga, I. (2019b). Lexico-syntactic interactions during the processing of temporally ambiguous L2 relative clauses: An eye-tracking study with intermediate and advanced Portuguese-English bilinguals. PLoS ONE 14(5):e0216779. https://doi.org/10.1371/journal.pone.0216779

Soares, A. P., Perea, M., & Comesaña, M. (2014b). Tracking the Emergence of the Consonant Bias in Visual-Word Recognition: Evidence with Developing Readers. *PLOS ONE, 9*(2), e88580. https://doi.org/10.1371/journal.pone.0088580

Soares, A. P., Velho, M., & Oliveira, H. M. (2020). The role of letter features on the consonant-bias effect: Evidence from masked priming. *Acta Psychologica, 210*. https://doi.org/10.1016/j.actpsy.2020.103171

Tainturier, M.-J. (2019). A theory of bilingual spelling in alphabetic systems. In T. Olive & C. Perret (Eds.), *Spelling and Writing Words: Theoretical and methodological advances (Studies in Writing; Vol. 39)*. Brill. https://doi.org/10.1163/9789004394988

Text2Phonetics. (n.d.). *Text2Phonetics*. Retrieved July 01, 2020, from http://www.photransedit.com/Online/Text2Phonetics.aspx

Titone, D. A., & Libben, M. (2014). Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: A cross-modal priming investigation. *The Mental Lexicon, 9*(3), 473–496. https://doi.org/10.1075/ml.9.3.05tit

Tophonetics. (n.d.). Retrieved July 01, 2020 from https://tophonetics.com/

Valente, D., Ferré, P., Soares, A. P., Rato, A., & Comesaña, M. (2017). Does phonological overlap of cognate words modulate cognate acquisition and processing in developing and skilled readers? *Language Acquisition, 25*(4). https://doi.org/10.1080/10489223.2017.1395029

van Hell, J. G., & Dijkstra, T. (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychonomic Bulletin & Review, 9*, 780–789. https://doi.org/10.3758/BF03196335

van Heuven, W. J. B., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language, 39*(3), 458–483. https://doi.org/10.1006/jmla.1998.2584

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory and Cognition, 15*, 181–198. https://doi.org/10.3758/BF03197716

Voga, M., & Grainger, J. (2007). Cognate Status and Cross-script Translation Priming. *Memory and Cognition, 35*(5), 938–952. https://doi.org/10.3758/bf03193467

Wells, J. C. (1982). Accents of English. Volume 1: An Introduction (pp. i–xx, 1–278), Volume 2: The British Isles (pp. i–xx, 279–466), Volume 3: Beyond the British Isles (pp. i–xx, 467–674). Cambridge University Press.

Williams, C. B. (1937). The use of logarithms in the interpretation of certain entomological problems. *Annals of Applied Biology, 24*, 404–414.

Open Practices Statement The materials reported in this manuscript are freely available for use at https://www.psi.uminho.pt//pt/CIPsi/Laboratorios_Investigacao/Psicolinguistica/Documents/PHOR_in_One_LDB.zip and in the supplementary materials.