Behavior Research Methods, *in press*

ESCOLEX: A grade-level lexical database from European Portuguese Elementary to Middle School textbooks

Ana Paula Soares[1], José Carlos Medeiros[2], Alberto Simões[3,4], João Machado[1], Ana Costa[1], Álvaro Iriarte[3], José João de Almeida[4], Ana P. Pinheiro[5], & Montserrat Comesaña[1]

[1] Human Cognition Lab, CIPsi, School of Psychology, University of Minho, Portugal.

[2]Porto Editora, Portugal.

[3]Centre for Humanistic Studies, University of Minho, Portugal.

[4]Computer Science and Technology Center, University of Minho, Portugal.

[5]Neuropsychophysiology Lab, CIPsi, School of Psychology, University of Minho, Portugal.

Corresponding author:

Ana Paula Soares

Human Cognition Lab, CIPsi, School of Psychology,

University of Minho

Campus de Gualtar

4710-057 Braga, Portugal

E-mail: asoares@psi.uminho.pt

Phone: + 351 253604236

Abstract

In this paper we introduce ESCOLEX, the first European Portuguese children's lexical database with grade-level-adjusted word frequency statistics. Computed from a 3.2 million word corpus, ESCOLEX provides 48,381 wordforms extracted from 171 Elementary and Middle School textbooks for 6 to 11 year-old children attending the first six grades in the Portuguese educational system. Similarly to other children's grade-level databases (e.g., Carroll, Davies, & Richman, 1971; Corral, Ferrero, & Goikoetxea, 2009; Lété, Sprenger-Charolles, & Colé, 2004; Zeno, Ivenz, Millard, & Duvvuri, 1995), ESCOLEX provides four frequency indices for each grade: overall word frequency (*F*), index of dispersion across the selected textbooks (*D*), estimated frequency per million words (*U*), and standard frequency index (*SFI*). It also provides the new measure of contextual diversity (*CD*). Additionally, the number of letters in the word, part-of-speech, number of syllables, syllable structure, and adult frequencies taken from P-PAL (a European Portuguese corpus-based lexical database - Soares et al., 2010; *in press*) are also provided. ESCOLEX is a useful tool for both researchers interested in language processing and development, and professionals in need of verbal material adjusted to children's developmental stages. ESCOLEX can be downloaded at http://brm.psychonomic-journals.org/content/supplemental or at http://p-pal.di.uminho.pt/about/databases.

Key words: children lexical databases, word frequency, child language processing, reading, literacy.

Running Head: ESCOLEX: A Portuguese grade-level lexical database

Introduction

In the last decade, a developmental focus has gained strength within experimental psycholinguistic research. As a consequence, in addition to studies targeting adult individuals who have already acquired language abilities or adults with some kind of linguistic impairment (e.g., aphasia, alexia, acquired dyslexia), young participants have been increasingly included in more recent studies with the aim of understanding how language knowledge is represented in the human mind and how language is understood and produced in different developmental stages.

Beyond the traditional interest in areas such as language acquisition (e.g., Bennedict, 1979; Bloom, 1973; Brown, 1973; Goldfield & Reznick, 1990; Slobin, 1973; Tomasello, 2003), reading and writing (e.g., Adams, 1990; Dickinson, & Snow, 1987; Ehri, 1995; Mason, 1980; Perfetti, 1985), and developmental language disorders (e.g., Castles & Coltheart, 1993; Lieberman & Shankweiler, 1985; Snowling, 1980), topics traditionally associated with adult research have increasingly raised the interest of researchers conducting experimental psycholinguistic studies with children. For instance, several studies have been recently developed with the aim of exploring which factors affect word recognition in children at different developmental stages and how the influence of those factors changes over time (e.g., Castles, Davis, Cavalot, & Forster, 2008; Lété, Peereman, & Fayol, 2008; Newman, 2012; Pattamadilok, Morais, de Vylder, Ventura, & Kolinsky, 2009; Perea, Soares, & Comesaña, *in press*; Perea, Panadero, Moret-Tatay, & Gómez, 2012; Ventura, Kolinsky, Pattamadilok, & Morais, 2008).

The increased inclusion of children in experimental psycholinguistic research answers the need to overcome several limitations in studies with adults. For example, children's studies allow for the investigation of cognitive processes that are often difficult to disentangle in adult populations in which language abilities have already been acquired and the lexicon is fully developed. Another advantage is the possibility of studying processes and mechanisms that may be more salient during specific language acquisition stages, thus avoiding not only the occurrence of ceiling effects, but also potential interaction effects that may mask the influence of those processes on linguistic performance over time. A clear example

of these advantages is the study of the influence of phonology on visual word recognition. This influence may be better captured in early stages of reading development, since access to phonological codes weakens as age or reading ability increases (e.g., Doctor & Coltheart, 1980; Newman, 2012; Turkeltaub, Gareau, Flowers, Zeffiro, & Eden, 2003; Unsworth & Pexman, 2003).

In fact, even though phonology and orthography are automatic processes that are highly interconnected in early stages of visual word recognition (see Rastle & Brysbaert, 2006 for a review), accumulating evidence suggests that orthographic processing drives word recognition in skilled mature readers, even during speech processing. Seidenberg and Tanenhaus (1979) showed that participants make faster rhyming decisions when pairs of aurally presented words share spelling (e.g., toast–roast), as opposed to word pairs that do not (e.g., toast–ghost). Recent studies using different experimental tasks and paradigms (e.g., Chéreau, Gaskell, & Dumay, 2007; Damian & Bowers, 2009; Muneaux & Ziegler, 2004; Pattamadilok, Perre, Dufau & Ziegler, 2009; Peereman, Dufour & Burt, 2009; Rastles, McCormick, Bayliss & Davis, 2011; Taft, Castles, Davis, Lazendic, & Nguyen-Hoan, 2008; Ventura, Morais, & Kolinsky, 2006; Ziegler, Ferrand, & Montant, 2004) have systematically confirmed these results. However, in pre-readers that interference is not observed. For example, Ziegler and Muneaux (2007) showed that before literacy, auditory lexical decisions are not influenced by the spelling of words, but as soon as children have contact with words in printed materials (i.e., during literacy development), spoken word recognition becomes affected by the orthographic knowledge that children have then developed (see also Goswami, Ziegler, & Richardson, 2005; Ventura et al. 2006, 2008). These results suggest that children become sensitive to the orthographic constraints of language only after being exposed to written words (i.e., when they start to read and write). Thus, contrary to what is observed in adults, the study of pre-readers allows the investigation of "purer" phonological effects on lexical access.

The importance of phonology during this developmental stage is also demonstrated by a vast number of studies showing, on the one hand, that phonological processing deficits have a relevant impact on reading difficulties observed in dyslexic children in alphabetic languages (e.g., Harm & Seidenberg, 1999; Vellutino,

Fletcher, Snowling, & Scanlon, 2004; Ziegler & Gaswani, 2005) and, on the other hand, that pre-school phonological awareness programs have significant effects on the development of skills necessary for successful reading (e.g., Adams, 1990; Lonigan, Burgess, & Anthony, 2000). Recent neurophysiological evidence also supports these findings. As proposed by the dorsal–ventral neuroanatomical model of reading acquisition (Shaywitz et al., 2002), temporal-parietal areas and the inferior frontal cortex, associated with phonological processing, are more activated in younger readers. However, as reading competence develops, those regions become less activated and there is an increased engagement of inferior temporal and occipital regions, which are associated with orthographic processing (e.g., Blomert, 2011; Booth et al., 2004; Newman, 2012; Turkeltaub et al., 2003).

Together, these studies show that recruiting children as research participants for experimental psycholinguistic studies contributes to a deeper understanding of the dynamics of language representation and processing while these abilities are still developing, thus avoiding the intrinsic limitations of fully acquired language abilities and a fully developed lexicon. Additionally, at a theoretical level, results from child studies may also inform cognitive models of visual (e.g., Coltheart, Rastle, Perry, Langdon & Ziegler, 2001; Plaut, McClelland, Seidenberg & Patterson, 1996; Reichle, Pollatsek, Fisher, & Rayner, 1998) and spoken word recognition (e.g., Luce, Pisoni, & Goldinger, 1990; McClelland & Elman, 1986), and improve our understanding of the specific mechanisms that underlie typical and atypical pathways in language development, such as dyslexia.

The increasing interest in developing children's studies with an experimental focus requires, as in research with adults, the existence of lexical databases containing relevant psycholinguistic information, such as word frequency. Word frequency indexes the number of times a word occurs in a language and represents one of the most robust effects in the history of psycholinguistics (see Monsell, 1991 for a review). High-frequency words are processed more quickly and accurately than low-frequency words, which has been taken as evidence that the systems involved in language representation and processing are sensitive to the statistical properties of an individual's linguistic experience.

Nonetheless, given the emerging nature of children's lexicons, it is critical to determine whether frequency counts obtained from adult corpora (presumably representative of a fully developed lexicon) are compatible with those obtained from children's corpora. Some authors argue that the statistical regularities associated with emerging lexicons (such as phonological neighborhood density) inflate when adult norms are applied to children's studies because their lexicons are more limited, thus resulting in an overestimation of behavioral effects (see Dollaghan, 1994). However, other authors claim that because children's lexicons are more limited, using adult corpora may underestimate the impact of statistical variables on children's language learning (see Coady & Aslin, 2003).

Although these perspectives point towards opposite effects in the use of adult corpora in children's studies, both recognize that the use of adult corpora may bias results obtained from children's studies. In line with this concern, Zevin and Seidenberg (2002) recommend the use of database counts based on children's corpora instead of adult counts. Studying the age-of-acquisition (AoA) effect, the authors found that the *Educator's Word Frequency Guide* from Zeno, Ivenz, Millard, and Duvvuri (1995) was more closely correlated with word reading latencies of children than earlier adult counts, such as those obtained by Kŭcera and Francis (1967). According to Zevin and Seidenberg (2002), these results were due to the fact that Zeno et al.'s frequency norms were computed from children-targeted texts. These results showed that using norms adjusted to children's developmental stages is critical to obtain reliable effects.

Another controversy has to do with the sources these norms are obtained from. The typical strategy for adult frequency norms is to compile large amounts of text produced by adults, i.e., corpora (see Soares et al., *in press*, for a recent example). However, the fact that children's language production ability lags considerably behind their ability to understand language raises serious theoretical and methodological problems to the development of norms based on children's productive lexicons (see Smolensky, 1996). To overcome this problem several databases have been created using corpora designed specifically for children, such as storybooks or schoolbooks.

This strategy has been widely acknowledged and many international grade-level databases have been developed by assessing how often words are experienced by children during their exposure to printed language. The existing databases include the *American Heritage Word Frequency* (Carroll, Davis & Richman, 1971) that provides frequencies for 86,741 American English words extracted from a 5.09 million word corpus with publications widely read by American 7-15 year-old children, and the aforementioned *Educator's Word Frequency Guide* from Zeno et al. (1995) with 154,941 entries from a corpus of 17 million words used from pre-school to college. Recently, other databases have been developed in different languages such as the *Children's Printed Word Database* (CPWD: Stuart, Dixon, Masterson, & Gray, 2003 - for an extension, see Masterson, Stuart, Dixon, & Lovejoy, 2010) for British English; the NOVLEX (Lambert & Chesnet, 2001) and the MANULEX (Lété, Sprenger-Charolles, & Colé, 2004 - for an extension, see Peereman, Lété, & Sprenger-Charolles, 2007) for French; the LEXIN (Corral, Ferrero, & Goikoetxea, 2009) and the ONESC (Martínez & García, 2008) for Spanish; and the *Lessico Elementare* (Marconi, Ott, Pesenti, Ratti, & Tavella, 1993) for Italian. However, children's frequency counts are still unavailable for European Portuguese (see Note 1).

ESCOLEX was developed to overcome this gap. Computed from a 3.2 million-word corpus developed from 171 elementary and middle school textbooks, ESCOLEX is the first European Portuguese children's lexical database to provide grade-level word frequency statistics for 48,381 wordforms. Word frequencies are provided for 6 to 11 year-old children attending the first six grades in the Portuguese educational system. In accordance with other grade-level children's databases (e.g., Carroll et al., 1971; Corral et al., 2009; Lété et al., 2004), ESCOLEX provides four word frequency statistics (overall word frequency – *F*; index of dispersion across the selected textbooks – *D*; estimated frequency per million words – *U*; and standard frequency index -*SFI*), as well as five other psycholinguistic indices (number of letters in the word, parts-of-speech, number of syllables, syllable structure, and adult frequencies extracted from P-PAL, a European Portuguese corpus-based lexical database - see Soares et al., 2010; *in press*). ESCOLEX also provides a new measure of contextual diversity (*CD*), which indexes the number of contexts (i.e.,

different textbooks) in which a word appears, since recent evidence showed that CD is a relevant factor in word recognition both in adults (e.g., Adelman, Brown, & Quesada, 2006) and in developing readers (e.g., Perea, Soares, & Comesaña, in press). Indeed, in a recent study that aimed at testing to what extent *CD* and word frequency account for the lexical decision times of Portuguese young readers (fourth Grade), Perea et al. (*in press*) found that *CD* (and not word frequency) was the main determinant of word identification times, thus generalizing the data from Adelman et al. (2006) with college participants to fourth Grade children.

ESCOLEX is a useful tool for both researchers interested in language processing and development both in typical and atypical populations, and also in other cognitive and neuroscience areas (e.g., memory), as well as for professionals in need of verbal material adjusted to children's developmental stages. The use of this database may contribute not only to a more appropriate selection of stimuli to be used in experimental research but also to the development of teaching programs including words adjusted to the educational and age levels of their targets. ESCOLEX may also contribute to the development of intervention programs for children with speech and/or writing difficulties, and to the creation of sensitive measures of reading ability, that may avoid the floor effects that are usually observed in standardized tests (see Bowey, 2005). Lastly, it is worth noting that ESCOLEX may additionally represent a useful tool for children-targeted book publishers and writers by allowing the adjustment of vocabulary according to the age and/or educational level of their readers. ESCOLEX can be downloaded at http://brm.psychonomic-journals.org/content/supplemental or at http://p-pal.di.uminho.pt/about/databases.

Material and methods

Corpus sampling

ESCOLEX was developed from a corpus of 171 European Portuguese Elementary and Middle School textbooks published between 1998 and 2007 and provided by Porto Editora, one of the major publishers in Portugal. The textbooks in ESCOLEX were used in the formal training of 6- to 11-year-old children attending the first six grades of the official Portuguese educational system.

In Portugal, schooling starts at 6 years old when children enter Elementary School (hereafter also First Cycle), which comprises the first four grades (first Grade: 6-year-olds; second Grade: 7-year-olds; third Grade: 8-year-olds; and fourth Grade: 9-year-olds; hereafter G_1, G_2, G_3, G_4, correspondingly). The curriculum is compulsory, equal for all students, and covers four different subjects, namely Portuguese, Mathematics, Environment Studies and Physical and Artistic Expression. Subsequently students enter Middle School (hereafter also Second Cycle), which comprises the fifth and sixth grades (fifth Grade: 10-year-olds and sixth Grade: 11-year-olds, hereafter G_5 and G_6, correspondingly). In this stage, eight compulsory subjects comprise students' educational *curricula*: Portuguese, Foreign Language (usually English, French or Spanish), Portuguese History and Geography, Mathematics, Natural Sciences, Visual and Technological Education, Musical Education and Physical Education. Usually at age 12 students enter the Third Cycle (which comprises grades 7 to 9) and then the Secondary Education (which comprises grades 10 to 12), before entering Higher Education.

The 171 textbooks included in ESCOLEX encompass the four Elementary school subjects (G_1-G_4) and seven of the eight Middle school subjects (G_5-G_6). Because ESCOLEX aims to be a close representation of the European Portuguese printed vocabulary to which children are exposed, the foreign language textbooks have been excluded. There are however loanwords in the ESCOLEX wordlist (e.g., 'cowboys', 'sandwich', 'show', 'rally', 'bowling', 'windsurf') since they have been adopted in European Portuguese and are now a part of the Portuguese vocabulary.

Figure 1 presents the distribution of the 171 textbooks in ESCOLEX according to subject for each grade (G_1 to G_6), educational stage (G_1-G_4 and G_5-G_6) and for all grades combined (G_1-G_6). In ESCOLEX one book corresponds specifically to one International Standard Book Number (ISBN).

<INSERT FIGURE 1 ABOUT HERE>

As shown in Figure 1, textbook distribution according to grade is fairly balanced, with an average of 27.5 Elementary School books (M_{G1} = 27; M_{G2} = 25; M_{G3} = 36; and M_{G4} = 22) and 38.5 Middle School books (M_{G5} = 37; and M_{G6}= 40). In each grade, textbook distribution according to subject (considering the total number of books in each educational level) indicates a higher percentage of Portuguese textbooks in Elementary School (51.8%), followed by Mathematics (25.5%), Environment Studies (14.2%) and Physical and Artistic Expressions (8.5%). As for Middle School books, the distribution across the seven subjects is more balanced, with a slight predominance of Mathematics (21.5%), followed by Natural Sciences (20.0%), Portuguese (16.9%), Portuguese History and Geography (16.9%), Musical Education (10.8%), Visual and Technological Education (7.7%) and, lastly, Physical Education (6.2%). Considering all grades (G_1-G_6), Portuguese books (38.6%) outnumber Mathematics (24.0%), Environment Studies (8.8%), Natural Sciences (7.7%), Portuguese History and Geography (6.4%), Plastic Expression and Education (5.4%), Musical Education (4.1%), Visual and Technological Education (2.6%) and Physical Education (2.4%). Four textbooks have been included as part of both G_3 and G_4 and 12 textbooks as part of both G_5 and G_6, since they are recommended for both grades of either educational stage. In this case, even though the textbooks were accounted for in both G_3 and G_4 or G_5 and G_6 as specified above, they were only counted once in the data featuring in the Elementary (G_1-G_4) and Middle School (G_5-G_6) educational stages, as well as in all ESCOLEX grades (G_1-G_6).

The ESCOLEX corpus was compiled from the PDF versions of the original 171 textbooks. All textbook areas were indexed, except headers, prefaces, introductory notes, bibliography and references to complementary materials presented at the end of some textbooks. These sections were excluded because they are typically addressed to teachers and/or parents and the vocabulary does not reflect children's lexicons. The Porto Editora tagger was then used in order to recognize words (i.e., set of letters between two blank spaces). Hyphenated words, numerals, and proper nouns with the same orthography as common nouns were indexed.

The original 171 textbooks yielded a total corpus of 3,211,805 tokens. Nine data sheets were then compiled: six developed by assigning textbooks to their corresponding grade level (G_1, G_2, G_3, G_4, G_5, and G_6), and two by combining the data obtained into two educational stages (G_1-G_4 and G_5-G_6). The remaining sheet contains data for grades G_1 to G_6. Figure 2 shows the distribution of occurrences (tokens) in each ESCOLEX subcorpus according to grade and educational level considering the total corpus.

<INSERT FIGURE 2 ABOUT HERE>

The Middle School subcorpus (G_5-G_6) is the largest one (1,922,082 *vs*. 1,289,723 occurrences in Elementary School textbooks – G_1-G_4). In the Elementary School subcorpus, G_3 and G_1 present the largest (525,041 tokens corresponding to 15% of the overall corpus and to 39% of the Elementary School subcorpus) and the lowest number of tokens (193,548 corresponding to 5% of all corpus and 14% of the Elementary School subcorpus). As mentioned before, this is due to the fact that more books have been integrated in G_3 than in any other grade (cf. Fig. 1), and to the fact that textbooks for beginners as included in G_1 are typically smaller and contain fewer words. In the Middle School subcorpus the distribution is more balanced: G_5 has 1,032,449 occurrences (corresponding to 29% of the entire corpus and 46% of the Middle School subcorpus) and G_6 has 1,185,435 occurrences (corresponding to 33% of the entire corpus and 53% of the Middle School subcorpus).

In order to define the lexical entries (wordforms) in each of the nine ESCOLEX data sheets, we used a similar strategy to the one used for P-PAL (for details see Soares et al., *in press*). Proper nouns and isolated syllables (common in G_1 books), abbreviations (e.g., *vol.* [for the English word 'volume'] or *art.* [for the English word 'article']), symbols and unconventional orthographic forms (e.g., @ or €) were eliminated by cross-checking the lexical entries in ESCOLEX both with the morphologic analyzer *JSpell* (Simões & Almeida, 2001) and with the lexical entries in P-PAL. Numerals, loanwords, and proper nouns with the same orthography as common nouns (e.g., the adjective *clara* [light] is also a proper noun) were maintained.

Hyphenated words were also kept, except verbs with clitic pronouns. These inflected verb forms are a combination of two or more words in a compound verb form and were therefore split into their constituents. For example, the verb form *preparavam-se* [they prepared themselves] was split into the verb form *preparavam* [they prepared] and the clitic pronoun *se* [themselves]. The original frequency of the compound verb form was added to the final verb form and clitic pronoun. Similarly, contractions were split into their lexical constituents (e.g., *dele* [his] is a contraction of the preposition *de* [of] and the personal pronoun *ele* [him] and was split into *de* and *ele*) and their original frequencies were assigned to each lexical item. Multiword items, i.e., unhyphenated words such as phrases, idioms and collocations, were also split into their lexical constituents and the original frequency value was added to each item. It is worth noting that in ESCOLEX there is one single entry for nonhomophonic homographs (e.g., *sede* ['sedə], the Portuguese word for 'thirst'; and *sede* ['sɛdə], the Portuguese word for 'headquarters') and homonyms (e.g., *castanha* [noun], the Portuguese word for 'chestnut'; and *castanha* [adjective], the feminine for 'brown').

Based on this procedure, the total lexicon in ESCOLEX comprises 48,381 different wordforms (types). Even though this number is inferior to the number of forms contained in the database of Carroll et al. (1971) (that includes 86,741 forms), it is equivalent to the number of forms presented in MANULEX (with 48,886 wordforms), and superior to the number of forms presented in LEXIN (with 13,184 wordforms) or in the CPWD, both in the version published by Suart et al. (2003) (9,748 words) and in the updated version published by Masterson et al. (2010) (12,193 words). From the total number of wordforms (cf. Table 1), 8,316 words pertain to G_1 (from a subcorpus of 193,548 tokens), 13,019 to G_2 (from a subcorpus of 258,470 tokens), 20,444 to G_3 (from a subcorpus of 525,041 tokens), 19,486 to G_4 (from a subcorpus of 382,355 tokens), 31,429 to G_5 (from a subcorpus of 1,032,449 tokens) and 35,216 to G_6 (from a subcorpus of 1,194,284 tokens). Grades G_1-G_4 and G_5-G_6 contain 29,013 (from a subcorpus of 1,289,723 tokens) and 41,952 words (from a subcorpus of 1,922,082 tokens), correspondingly.

Results and discussion

The ESCOLEX database can be downloaded at http://brm.psychonomic-journals.org/content/supplemental or at http://p-pal.di.uminho.pt/about/databases as an excel file. This file contains nine spreadsheets, one corresponding to the total ESCOLEX corpus (G_1-G_6), six corresponding to each grade (G_1, G_2, G_3, G_4, G_5, and G_6), and the last ones corresponding to each educational level (G_1-G_4, and G_5-G_6). In each spreadsheet, 10 columns follow the lexical entries (wordforms), five of which provide information regarding children word frequency indices (overall word frequency – *F*; index of dispersion across the selected textbooks – *D*; estimated frequency per million words – *U*; standard frequency index – *SFI*; and the number of books in which the word appeared - *CD*), and the remaining five columns contain information about the grammatical and sublexical properties of words (number of letters in the word, part-of-speech, number of syllables, syllable structure), and adult word frequencies, obtained from the P-PAL lexical database (available at http://p-pal.di.uminho.pt/tools). The frequency indices were computed according to the methods adopted by Carroll et al. (1971), Lété et al. (2004), and Corral et al. (2009), and described by Breland (1996) as follows:

Frequency (*F*): number of times a word appears in the ESCOLEX corpus. *F* is given for each grade, for each educational stage (G_1-G_4 and G_5-G_6), and for the total corpus (G_1-G_6). For the total corpus, *F* varies between 1 (13,118 words) and 158,353 tokens (the word occurring more frequently is the function word *a*, the Portuguese feminine word for 'the').

Dispersion (*D*): distribution of the frequency of each word across textbooks in each ESCOLEX grade level. *D* is also computed for each grade separately, for G_1-G_4, G_5-G_6 and G_1-G_6. *D* ranges from 0 to 1: $D = 0$ when all occurrences of the word are found in a single textbook regardless of its frequency, and $D = 1$ when the frequencies are distributed exactly in the same proportion across textbooks. The formula for calculating *D* is:

$$D = [\log(\Sigma\, p_i) - [(\Sigma\, p_i \log p_i) / \Sigma\, p_i\,]] / \log(n),$$

where *n* is the number of textbooks in each subcorpus ($n = 27$ in G_1, 25 in G_2, 36 in G_3, 22 in G_4, 37 in G_5, 40 in G_6, 106 in G_1-G_4, 65 in G_5-G_6, and 171 in G_1- G_6); *i* is the textbook number (1, 2, . . . , *n*), and p_i

is the frequency of a word in the i^{th} textbook, with $p_i \log p_i = 0$, if $p_i = 0$. In the total corpus, 15,832 words have a *D* value of 0 and two words (the determiner *o* [the masculine form of 'the'] and the preposition *com* [with]) have a *D* value of 0.96.

Estimated frequency per million words (*U*): estimated frequency value of each word adjusted to *D*. When $D = 1$, *U* is computed as the frequency *per* million words. When $D < 1$, the value of *U* is adjusted downward. When $D = 0$, *U* has a minimum value based on the mean weighted probability of the word's occurrence across the textbooks. The adjustment is made using the following formula:

$$U = (1{,}000{,}000 / N)\,[FD + (1 - D) * f_{\min}],$$

where *N* is the total number of words in each ESCOLEX grade and educational stage (191,942 in G_1, 257,950 in G_2, 520,488 in G_3, 376,732 in G_4, 1,025,703 in G_5, 1,201,480 in G_6, 1,277,611 in G_1-G_4, 1,906,284 in G_5-G_6, and 3,183,895 in G_1- G_6), *F* is the frequency of the word in each subcorpus, *D* is the index of dispersion, and $f_{\min}$ is $1/N$ times the sum of the products of f_i and s_i, where f_i is the frequency in textbook *i*, and s_i is the number of words in that textbook. We followed Lété et al. (2004) and considered that *U* is a better measure of word frequency, since it allows more reliable and direct comparisons between word frequencies across subcorpora. For the total corpus, *U* varies between a minimum of 0.0001 (for the loanword bowling) and a maximum of 47,049.33 (for the function word *a* [the feminine form for *the*]).

Standard frequency index (*SFI*): an index derived directly from *U* (thus assuming some of its properties) that may be used as a simple and convenient way of indicating frequency counts. For example, a wordform where *SFI* = 90 is expected to occur once in every 10 words, one where *SFI* = 80 is expected to occur once in every 100 words, one where *SFI* = 70 is expected to occur once in every 1000 words, and so on. A reference value is an *SFI* of 40, which means a word occurs once in a million words. *SFI* is computed using the following formula:

$$SFI = 10 * [\log 10\,(U) + 4].$$

For example, the wordforms *diário* [diary] and *aluno* [student] have the same frequency in G_1 (46 occurrences). However, they have an estimated frequency (*U*) of 15.03 and 121.32 *per* million words and

different *D* values (0 and 0.49 each), because the word *diário* was found 46 times in the same textbook, and the word *aluno* was found in seven different textbooks. Hence, their corresponding *SFI* values are 51.77 and 60.84 (the word *diário* occurs approximately 10 times in a million words, and *aluno* occurs approximately 100 times in a million words). This shows that even if words have the same number of occurrences in the corpus as described above, they may differ in the remaining frequency indices, which should be taken into account for a more detailed view of the frequency distribution of words in children's lexicons, as opposed to the clustered view offered by classic frequency measures. For the total ESCOLEX corpus, *SFI* varies between a minimum of 0.77 (again for the loanword bowling) and a maximum of 86.73 (for the function word *a* [the feminine form for *the*]).

Contextual Diversity (CD): number of textbooks in which the word appears divided by the total number of textbooks considered in each ESCOLEX grade level. *CD* ranges from 0 to 1: *CD* = 1 when the occurrences of the word are found in all textbooks in each subcorpus (G_1 = 27; G_2 = 25; G_3 = 36; G_4 = 22; G_5 = 37; G_6 = 40; G_1 _ G_4 = 106; G_5 _ G_6 = 65; and G_1 _ G_6= 171). In the total corpus, *CD* ranges between 0.01 (15,832 words) and 1 (22 function words with the exception of the verb form *faz* [does]). Of note, although CD and D are two conceptually similar measures (both capturing the variety of contexts – books – in which a word appears), the CD measure is preferable to D if researchers are interested in a word diversity measure that is independent of the word frequency distribution (token frequency) in those contexts (as it is indexed by the D measure). Albeit related, these indices differ because D introduces a different component to CD, as it adjusts it according to the proportion of the word frequency distribution. It is therefore a more refined measure (note that in ESCOLEX there are no words with a D value of 1, but several with a CD value of 1). For instance, the wordform *varanda* [balcony] occurs three times and the wordform *desaparecida* [feminine for missing] occurs six times in the G_1 corpus. Both words occur in three different contexts (books) in a total of 27 possible books, which is why both have a value of CD = .11. However, the wordform *varanda* [balcony] has a D value of .33 (1 + 1 + 1), while the wordform *desaparecida* [feminine for missing] has a D value of .31 (3 + 1 + 2). Evidently these two measures are empirically related (the *Pearson* correlation

between them is .75, $p < .001$ in the total corpus), but they are not redundant. Including them both in the ESCOLEX database enhance the indices provided by the database and furnish researchers with the possibility of selecting the index that best suits their own research requirements.

Number of letters (*Nlett*): number of letters in each ESCOLEX wordform. In the total corpus, Nlett ranges between 1 (six words with one letter, namely *o* [the], *e* [and], *é* [is], *à* [to the], *a* [the] and *ó* [hey] – function words, verb forms and interjections) and 22 letters (the word *otorrinolaringologista* [ENT specialist]). The mean extension (i.e., number of letters) in ESCOLEX is 8.39 letters (*SD* = 2.38). Eight-letter words are the most frequent in ESCOLEX (8,480 words) and they account for 17.5% of the overall lexicon. As stated by Soares et al. (*in press*), these data suggest that Portuguese is a synthetic language with a rich morphology. New words and word forms can be put together either by adding prefixes and suffixes, as in *des-entendi-mentos* [disagreements] (noun formed by derivation) or *cant-á-va-mos* [we used to sing] (inflected verb form), or by combining words or root words formed by composition such as *malmequer* [daisy] or into hyphenated words (e.g., *surdo-mudo* [deaf-mute]).

Number of syllables (*Nsyll*): number of syllables in each ESCOLEX wordform. For the total corpus, Nsyll ranges between 1 (440 words) and 10 syllables (four words, three of which are hyphenated: otorrinolaringologista [ENT specialist], científico-pedagógico [scientific and pedagogical], contra-revolucionário [counterrevolutionary] and topográfico-geológica [topographic and geological]). Three and four-syllable words are the most frequent syllable type (16,635 and 15,148 words, which represents 34.4% and 31.3% of the total ESCOLEX corpus, correspondingly), followed by words with two (7,242 wordforms, 15% of the total corpus) and five syllables (6,799 wordforms, 14.1% of the total corpus). Nine and ten-syllable words are the least frequent type (five and four words each). One-syllable words, very frequent in languages like English (see Fenk-Oczlon & Fenk, 2008), only account for 0.9% of the overall ESCOLEX lexicon.

Syllable structure (Syllest): set of consonants (C) and vowels (V) forming the orthographic segment of each wordform. In the total corpus, there are 4,097 different syllable structures: CV-CV-CVC (e.g., the

word *caracol* [snail]) is the most frequent syllable structure in ESCOLEX (1,436 words), followed by CV-CV-CV (1,132 words – e.g., the word *número* [number]), CVC-CV-CVC (1,101 words – e.g., *possível* [possible]), and CV-CVC (1,027 words – e.g., the verb *fazer* [to do]). From the total number of different syllable structures, 62% are infrequent structures, occurring less than three times in the database (e.g., structure VC-CCVV-CVV-CV as in *empreiteiro* [contractor], or V-CVC-CCVVC as in *electrões* [electrons]). Five-syllable words have more diversified syllable structures (1,209 different structures representing 29.5% of all syllable structures in ESCOLEX), followed by four (1,113 words, 27.2% of the total), six (706 words, 17.2% of the total) and three-syllable words (567 words, 13.8% of the total). Words with one and ten syllables have the least diversified syllable structures in ESCOLEX (22 and four each). The orthographic boundaries of the wordforms are signaled with a hyphen (-). Stress is signaled with an apostrophe (‘).

P-PAL frequency (P-PALfreq): number of occurrences of each ESCOLEX word in the P-PAL wordform lexical database (Soares et al., 2010; *in press*). P-PAL is a free web application that computes a wide range of word-related statistics (e.g., word frequency, bigram, biphone and syllable frequency, orthographic and phonological structure, morphological information, grammatical class, orthographic and phonological similarity – available at http://p-pal.di.uminho.pt/about/project), based on adult corpora of over 227 million words (Soares et al., *in press*).

Part-of-Speech (*POS*): grammatical information of each ESCOLEX wordform. This information was obtained from P-PAL (http://p-pal.di.uminho.pt/tools). As in P-PAL, content and function words in ESCOLEX cover the following classes: nouns (N), adjectives (ADJ), verbs (V), adverbs (ADV), conjunctions (CONJ), determiners (DET), interjections (INT), quantifiers (QUANT), prepositions (PREP), and pronouns (PRON) (see Soares et al., *in press*, for details). However, because syntactic ambiguity is very common in the Portuguese lexicon, where words like *ilustrado* [illustrated] can be used both as a verb form and an adjective, we decided to include in ESCOLEX all the corresponding grammatical classes the word has been assigned to in P-PAL. POS tags are comma separated. Similar to other children’s lexical databases

(e.g., LEXIN, MANULEX), ESCOLEX shows that the vast majority of words (99.3%) in early reading vocabularies are content words (i.e., nouns, verbs, adjectives, adverbs and interjections) rather than function words (i.e., pronouns, determiners, quantifiers, prepositions and conjunctions). The most frequent grammatical classes in the total corpus are verbs (48.6%), nouns (34.9%), and adjectives (14.2%); the least frequent grammatical classes include interjections (0.11%), conjunctions (0.15%) and determiners (0.18%).

Table 1 presents word distribution information in each of the nine ESCOLEX grade levels according to the total number of wordforms (types), the number of words occurring only once (a phenomenon known as hapax legomena*),* and the number of words occurring five times or more.

<INSERT TABLE 1 ABOUT HERE>

As shown in Table 1, children's vocabulary in ESCOLEX increases significantly from grade to grade, except from G_3 to G_4. This result reflects the fact that fewer textbooks were included in G_4 (cf. Fig. 1), which has led to a smaller and less diversified subcorpus. Nonetheless, similar to other children's lexical databases (e.g., CPWD, LEXIN, MANULEX), ESCOLEX findings indicate that, as children advance in their education, they are exposed to a growing number of printed words. This may indicate that the mental lexicon develops proportionately as children progress in their education. This may also reflect a fast acquisition, since the bulk of vocabulary growth during infancy seems to occur mostly with reading rather than with speaking or direct teaching (see Nagy & Herman, 1987).

The most significant increase in vocabulary is observed from G_1 to G_2 (4,703 new words were incorporated which corresponds to a 56.6% growth) and from G_2 to G_3 (7,245 new words were incorporated which corresponds to a 57% growth). An exception is the transition from G_4 to G_5, which, for the reasons previously presented, suggests an inflated growth (11,943 new words were added which corresponds to a 61.3% increase). When the educational stages are considered in the analysis, a significant increase in the vocabulary size is also observed. From Elementary (G_1-G_4) to Middle School (G_5-G_6) the vocabulary

increases 44.6% (12,939 new words) and from the first to the sixth-grade the vocabulary increases fourfold. Thus, as age and educational level increase, the number of words children are exposed to also increase exponentially.

If the observed increment in the number of words in textbooks promotes a great opportunity for children to develop a richer mental lexicon, the frequency with which they are exposed to words may constrain acquisition (Stuart et al., 2003). In fact, the distribution of hapax words (i.e., words occurring only once) in ESCOLEX indicates that approximately one third of the words children read in their textbooks consist of rare events. Even though this distribution may be inflated by the fact that ESCOLEX is based on wordform counts, it is similar to the distribution observed in other child databases based on wordform frequency counts (e.g., Carroll et al., 1971; Corral et. al., 2009; Stuart et al., 2003) – although the percentage of rare events in ESCOLEX is lower than the one in Carroll et al.'s (1971) and Stuart et al.'s (2003) corpora, where they account for approximately 50% of the lexicon. In databases with both lemma and wordform frequency counts – e.g., MANULEX (Lété et al., 2004) – the number of hapax words (23%) is lower for lemmas than for wordforms (31%), which is expected, since different hapax wordforms can pertain to one single lemma.

Considering that the number of encounters needed to learn a word varies between six to ten times (see Zahar, Cobb & Spada, 2001), it is possible that the number of hapax words in each ESCOLEX grade level (36% in G_1; 35% in G_2; 33% in G_3; 37% in G_4; 32% in G_5; 33% in G_6; 30% in G_1-G_4; 29.5% in G_5-G_6; and 27% in G_1-G_6) may interfere with efficient learning, since children are not sufficiently exposed to these words, as noted by Stuart et al. (2003). In fact, in light of systematic findings with adults (e.g., Balota et al., 2007; Zevin & Seidenberg, 2002), recent evidence with children has shown that word identification times are shorter when reading a word that occurs frequently in print than when reading an infrequent word (e.g., Moret-Tatay & Perea, 2011). Unsurprisingly, word frequency plays a central role in all current computational models of visual word recognition and reading (e.g., Coltheart et al., 2001; Plaut et al., 1996; Reichle et al., 1998), even though the mechanisms according to which word frequency is implemented vary

from model to model (e.g., see Norris, 2006 for a review). Additionally, recent findings have shown that word accessibility in the lexicon may not be determined exclusively by a mechanism of pure repetition (i.e., word frequency per se). The effect can be modulated by the number of contexts in which a word appears for both adults (e.g., Adelman et al., 2006) and children (Perea et al., in press). According to the recent study of Perea et al. (*in press*), the main determinant of word identification times in children is the number of contexts (textbooks) in which the word appears (i.e., contextual diversity) and not the number of times a word appears in a text (i.e., word frequency per se). Contextual diversity enhances the probability of word occurrence and fosters the development of richer lexical and semantic representations, as opposed to words appearing in more restrictive contexts. This may explain why in the study of Perea et al. words occurring in more contexts were recognized more quickly than words appearing in fewer contexts, even if they shared the same frequency of occurrence. In the example provided above, the words *diário* [diary] and *aluno* [student] have the same frequency of occurrence (46 times) but different *CD* values (1 and 7 respectively), which is why the word *aluno* is expected to be recognized faster and more accurately than the word *diário*. Therefore, the number of times children are exposed to printed words does not seem to suffice for effective learning: the diversity of situations and contexts of exposition seem to play a major role for successful word acquisition. Thus, and from an educational point of view, rather than providing children with repeated presentations of the same words in very specific or similar contexts, teachers should create learning opportunities so that the same word can be experienced in different contexts of occurrence. This will enhance the accessibility of children's lexical and semantic representations and consequently foster the development of their mental lexicon and literacy skills.

Another finding in ESCOLEX has to do with the fact that more than half of the words in the total corpus (54.5%) have low frequencies (below five tokens) which, in line with what has been pointed out for hapax words, can be somewhat inflated by the fact that ESCOLEX is a wordform database in which each inflectional word is distinguished (a lemma count would produce higher frequency values because the frequencies of each inflectional variant are summed). This percentage is even higher when grades (G_1 =

65.2%; G_2 = 66.2%; G_3 = 63.9%; G_4 = 67.8%; G_5 = 60.6%; and G_6 = 61.7%) and educational levels (G_1-G_4 = 59.3%; G_5-G_6 = 58.1%) are considered separately. This reveals that like other child databases based on wordform counts (e.g., Carroll et al., 1971; Corral et al., 2009; Stuart et al., 2003), ESCOLEX presents a strong bias toward low-frequency words, as expected by the Zipf law (1949) when applied to corpora.

Table 2 presents the mean, mode, and percentile values (P_{10}, P_{25}, P_{50}, P_{75} and P_{90}) for each of the five word frequency statistics provided by ESCOLEX (overall word frequency – *F*; index of dispersion across the selected textbooks – *D*; estimated frequency per million words – *U*; standard frequency index –*SFI*; and the number of books in which the word appeared - *CD*) in each of the nine grade levels.

<INSERT TABLE 2 ABOUT HERE>

Analysis of the distribution of word frequencies in each of the nine ESCOLEX grade levels indicates, as mentioned before, that the mean overall word frequency (*F*) is low in all grades. Words with frequency up to two ($F < 3$) account for about 50% of the lexicon between G_1 and G_4 (G_1 = 51.4%; G_2 = 51.9%; G_3 = 49.5%; G_4 = 53.6%) and words with a frequency of occurrence up to three ($F < 4$) account for more than 50% of the lexicon in G_5 (55.1%), G_6 (56.2%), G_1-G_4 (53.7%) and G_5-G_6 (52.5%). In the total corpus (G_1-G_6), 54.5% of the words have a frequency of occurrence up to four ($F < 5$). This high percentage of low-frequency words in ESCOLEX brings to light the controversy mentioned before regarding the number of encounters necessary for a printed word to be more effectively acquired by the child.

This bias towards low-frequency words is also observed in other frequency statistics provided by ESCOLEX. The estimated frequency (*U*) tends to decrease as grade increases, indicating an increase of vocabulary size and in the number of books incorporated in each grade level. Furthermore, considering a standard reference of 100 or more occurrences per million words for high-frequency words (as in the adult lexicon), it is clear that few words in ESCOLEX have that frequency value (G_1 = 741 words that represent 8.1% of the lexicon; G_2 = 848 words, 6.5% of the lexicon; G_3 = 927 words, 4.5% of the lexicon; G_4 = 896

words, 4.5% of the lexicon; G_5 = 842 words, 2.7% of the lexicon; G_6 = 855 words, 2.4% of the lexicon; G_1-G_4 = 913 words, 3.2% of the lexicon; G_5-G_6 = 864 words, 2.1% of the lexicon; and G_1-G_6 = 870 words, 1.8% of the lexicon). Even though the 100 most frequent words in each grade level represent a very small portion of the lexicon (G_1 = 1.2%; G_2 = 0.8%; G_3 = 0.5%; G_4 = 0.5%; G_5 = 0.3%; G_6 = 0.3%; G_1-G_4 = 0.4%; G_5-G_6 = 0.2%; and G_1-G_6 = 0.2%), analysis of these words reveals they account for a significant percentage of all tokens in each grade (G_1 = 67.4%; G_2 = 60.2%; G_3 = 56.9%; G_4 = 56.6%; G_5 = 56.3%; G_6 = 55.6%; G_1-G_4 = 57.4%; G_5-G_6 = 54.9%; and G_1-G_6 = 55.1%). The 500 most frequent words (6% of words in G_1; 3.8% in G_2; 2.5% in G_3; 2.6% in G_4; 1.6% in G_5; 1.4% in G_6; 1.7% in G_1-G_4; 1.2% in G_5-G_6; and 1% in G_1-G_6) account for 86%, 80.2%, 76.3%, 75.7%, 72.8%, 72.1%, 75.9%, 71.2% and 71.8% of all tokens, correspondingly. As in other children's databases (e.g., CPWD, LEXIN, MANULEX), the fact that such a reduced number of words represent a large part of the total frequency indicates that ESCOLEX has an irregular distribution of frequencies in each word set (i.e., a strong positively skewed distribution), which confirms in child corpora the fact that the frequency of any word is inversely proportional to its rank in the frequency table as expected by the Zipf law (1949).

Function words are the most frequent word types in ESCOLEX, as expected. In G_1-G_6, the set of the 100 most frequent words contains little more than 50 function words (54), but these words account for 44.5% of all tokens. The remaining 46 are content words, which account for only 9.6% of all tokens. This trend can be observed in all ESCOLEX grade levels, as content words represent a small percentage of the total frequencies. Nevertheless, as in the English (Stuart et al., 2003) and Spanish (Corral et al., 2009) children's databases, token frequency is inversely proportional to the percentage of content words, i.e., it decreases as the number of content words increases, until they come to represent 99% of the last 100 words of the first thousand.

This bias toward low-frequency words is also observed in the remaining ESCOLEX frequency statistics. For example, the dispersion index for the selected textbooks (*D*) is consistently low across grades, suggesting that on average words occur in less than 25% of the textbooks included for each grade level. This

finding is even clearer in the contextual diversity statistic (*CD*) results. Mean *CD* varies between .06 (G_1-G_6) and .16 (G_4), thus showing that words occur in average in 10 of the 171 books when considering the overall ESCOLEX corpus (G_1-G_6), and in 3.5 of the 22 books in G_4. The vast majority of words in ESCOLEX are therefore context-specific words. Along with the frequency distribution described above, this may inhibit word acquisition and the development of children's lexicons (Stuart et al., 2003), which may additionally compromise word recognition (Perea et al., in press). This is an issue that educators, teachers, editors and remaining professionals should be aware of.

Lastly, the *SFI* values, have an approximately symmetric distribution, with the mean close to the median at each grade. Consequently, as in MANULEX, the percentile values presented in Table 2 could be used as cutoffs for the selection of high and low-frequency words. As pointed out by Lété et al. (2004), mean *SFI* reflects the conceptual difficulty with which a word is learned by a child, with a decrease in the mean and mode values indicating an increase in word difficulty. According to Table 2, and similarly to other children's wordform databases, mean *SFI* values in ESCOLEX tend to decrease as grade level increases. The most significant drop takes place from G_4 to G_5, similarly to what was reported in MANULEX (although in MANULEX this drop takes place between G_3 and G_5 – for more details see Lété et al., 2004). It is possible that the transition to a different educational stage observed from G_4 to G_5 in the Portuguese educational system, and the associated critical changes in the school curricula (i.e., from four compulsory subjects in G_4 to eight in G_5), contribute to this finding.

Conclusion

ESCOLEX is the first European Portuguese child database to provide, similarly to other international databases (e.g., Carroll, et al., 1971; Stuart, et al., 2003; Corral et al., 2009; Lété, et al., 2004), several grade-level word frequency statistics as well as other word properties taken from the adult database P-PAL (a European Portuguese adult corpus-based lexical database - Soares et al., 2010, *in press*) for a total of 48,381

wordforms computed from a 3.2 million word corpus compiled from 171 Elementary and Middle School textbooks (6 to 11 year-old children).

Results obtained at each of the nine grade levels reveal that ESCOLEX may be a valuable and useful tool for research in language processing and development as well as in other cognitive (e.g., memory) and neuroscience areas. Consisting of a children's wordform frequency database with valuable psycholinguistic information, ESCOLEX allows for the selection of verbal stimuli adjusted to children's educational/developmental stages. ESCOLEX represents an additional resource to existing adult databases, as it focuses specifically on children's norms and may thus contribute to important issues raised by the use of frequency norms based on adult corpora.

Additionally, ESCOLEX may be a valuable resource for professionals working directly with children, specifically in terms of the development of teaching and intervention programs. For example, it can help in the development of programs for children with speech and/or writing difficulties, using materials adjusted to the educational stage and age of their targets and/or in the development of sensitive measures of reading ability. It is worth noting that ESCOLEX is also a useful tool for children's book publishers/writers by allowing them to adjust vocabulary according to the age and/or educational stage of their readers. ESCOLEX also provides valid suggestions on common vocabulary according to each educational stage in order to facilitate word acquisition. This is a particularly relevant aspect given the bias toward low frequency words found in ESCOLEX, which is also evident in other children's databases (e.g., CPWD, LEXIN, MANULEX).

Author note

This work is part of the research project “Procura-PALavras (P-PAL): A software program for deriving objective and subjective psycholinguistic indices for European Portuguese words” (PTDC/PSI-PCO/104679/2008), funded by FCT (Fundação para a Ciência e Tecnologia), and FEDER (Fundo Europeu de Desenvolvimento Regional) through the European programs QREN (Quadro de Referência Estratégico Nacional), and COMPETE (Programa Operacional Factores de Competitividade).

We are grateful to Porto Editora for providing the textbooks without which ESCOLEX would not have been possible.

Notes

1. Nonetheless, it is worth mentioning the existence of the unpublished PORTULEX database, which presents frequency and morphological information for 8,400 lemmas and 17,100 forms extracted from a corpus of 20 school textbooks (see http://www.fpce.up.pt/labfala/research.html).

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, Mass.: MIT Press.

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word naming and lexical decision times. *Psychological Science, 17*, 814–823.

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics*, *24*, 621–635

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., *et al.* (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459.

Benedict, H. (1979). Early lexical development: Comprehension and production. *Journal of Child Language*, *6*, 183-200.

Bloom, L. (1973). *One word at a time: The use of single word utterances before syntax*. The Hague: Morton.

Blomert, L. (2011). The neural signature of orthographic-phonological binding in successful and failing reading development. *NeuroImage*, *57*(*3*), 695-703.

Booth, J. R., Burman, D. D., Meyer, J. R., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2004). Development of brain mechanisms for processing orthographic and phonologic representations. *Journal Cognitive Neuroscience, 16*, 1234–49

Bowey, J. (2005). Grammatical sensitivity: Its origins and potential contribution to early reading skill. *Journal of Experimental Child Psychology, 90*, 318-343.

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

Brown, R. (1973). *A first language: The early stages*. London: George Allen & Unwin.

Caravolas, M. (2006). Refining the psycholinguistic grain size theory: effects of phonotactics and word formation on the availability of phonemes to preliterate children. *Developmental Science, 9*(5), 445– 447.

Carroll, J. B., Davies, P., & Richman, B. (Eds.) (1971). *The American Heritage word-frequency book*. Boston: Houghton Mifflin.

Castles, A., & Coltheart, M. (1993). Varieties of developmental dyslexia. *Cognition*, *47*, 149-180.

Castles A., Davis C., Cavalot P., & Forster K. (2008). Tracking the acquisition of orthographic skills in developing readers: Masked priming effects. *Journal of Experimental Child Psychology*, *97*, 165–182.

Chéreau, C., Gaskell, M. G., & Dumay, N. (2007). Reading spoken words: Orthographic effects in auditory priming. *Cognition*, 102, 341–360.

Coady, J. A., & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child Language, 30,* 441-469

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–256.

Corral, S., Ferrero, M., & Goikoetxea, E. (2009). LEXIN: A lexicaldatabase from Spanish kindergarten and first-grade readers. *Behavior Research Methods, 41*, 1009-1017.

Damian, M. F., & Bowers, J. S. (2009). Orthographic effects in rhyme monitoring: Are they automatic? *European Journal of Cognitive Psychology*, 22, 1–11.

Dickinson, D.K., & Snow, C.E. (1987). Interrelationships among prereading and oral language skills in kindergartners from two social classes. *Early Childhood Research Quarterly, 2*, 1-25.

Doctor, E. A., & Coltheart, M. (1980). Children's use of phonological encoding when reading for meaning. *Memory and Cognition*, *8*, 195–209.

Dollaghan, C. A., (1994). Children' s phonological neighbourhoods: Half empty or half full? *Journal of Child Language*, *21*, 257-272.

Ehri, L.C. (1995). Phases of development in learning to read words by sight. *Journal of Research in Reading*, *18*, 116-125.

Fenk-Oczlon, G., & Fenk, A. (2008). Complexity trade-offs between the subsystems of language. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), *Language complexity: Typology, contact, change* (pp. 43-65). Amsterdam: John Benjamins.

Goldfield, B.A., & Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of Child Language*, *17*, 171–183.

Goswami,U., Ziegler,J. C., & Richardson, U.(2005).The effects of spelling consistency on phonological awareness: A comparison of English and German. *Journal Experimental Child Psychology, 92*, 345–365.

Harm, M. W, & Seidenberg, M. S. (1999). Phonological, reading acquisition and dyslexia: Insights from connectionist models. *Psychological Review*, *106*(*3*), 491-528.

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lambert, E., & Chesnet, D. (2001). NOVLEX: Une base de données lexicales pour les élèves de primaire. *L'Année Psychologique*, *101*, 277-288.

Lieberman, I.Y., & Shankweiler, D. (1985). Phonology and the problems of learning to read and write. *Remedial and Special Education*, *6*, 8-17.

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. *Developmental Psychology, 36,* 596–613.

Luce, P.A., Pisoni, D.B., & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 122–147). Cambridge, MA: MIT Press.

Lété, B., Peereman, R., & Fayol, M. (2008). Consistency and word-frequency effects on spelling among first- to fifth-grade French children: A regression-based study. *Journal of Memory and Language*, *58*, 952-977

Lété, B., Sprenger-Charolles, L., & Cole, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, *36*, 156-166.

Marconi, L., Ott, M., Pesenti, E., Ratti, D., & Tavella, M. (1993). *Lessico Elementare. Dati statistici sull'italiano letto e scritto dai bambini delle elementari*. Bologna: Zanichelli.

Martínez, J. A., & Garcia, M. E. (2008). ONESC: A database of orthographic neighbors for Spanish read by children. *Behavior Research Methods*, *40*, 191-197.

Mason, J. (1980). When do children begin to read: an exploration of four year old children's letter and word reading competencies. *Reading Research Quarterly*, *15*, 203-227.

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2010). Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology*, *101*, 221-242.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G.W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 148-197). Hillsdale NJ: Erlbaum.

Moret-Tatay, C., & Perea, M. (2011). Is the go/no-go lexical decision task preferable to the yes/no task with developing readers? *Journal of Experimental Child Psychology*, *110*, 125–132.

Muneaux, M., & Ziegler, J. C. (2004). Locus of orthographic effects in spoken word recognition: Novel insights from the neighbour generation task. *Language and Cognitive Processes*, *19*, 641–660.

Nagy, W. E., & Herman, P.A. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In M. McKeown & M. Curtis (Eds.), *The Nature of Vocabulary Acquisition* (pp. 19-35). Hillsdale, NJ: Erlbaum Associates.

Newman, S. D. (2012). The homophone effect during visual word recognition in children: An fMRI study. *Psychological Research, 76*(*3*), 280-291.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*, 327–357.

Pattamadilok, C., Morais, J., de Vylder, O., Ventura, P., & Kolinsky, R. (2009). The orthographic consistency effect in the recognition of French spoken words: An early developmental shift from sublexical to lexical orthographic. *Applied Psycholinguist, 30*, 441–462.

Pattamadilok, C., Perre, L., Dufau, S., & Ziegler, J. C. (2009). On-line orthographic influences on spoken language in a semantic task. *Journal of Cognitive Neuroscience*, *21*(*1*), 169–179.

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

Perea, M., Panadero, V., Moret-Tatay, C., & Gómez, P. (2012). The effects of inter-letter spacing in visual-word recognition: Evidence with young normal readers and developmental dyslexics. *Learning and Instruction*, *22*, 420-430.

Peereman, R., Dufour, S., & Burt, J. S. (2009). Orthographic influences in spoken word recognition: The consistency effect in semantic and gender categorization tasks. *Psychonomic Bulletin and Review*, *16*(*2*), 363–368.

Peereman, R., Lété, B., & Sprenger-Charolles, L. (2007). Manulexinfra: Distributional characteristics of grapheme–phoneme mappings, and infralexical and lexical units in child-directed written material. *Behavior Research Methods*, *39*, 579-589.

Perfetti, C.A. (1985). *Reading ability*. New York: Oxford University Press.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi regular domains. *Psychological Review*, *103*, 56–115.

Rastle, K., & Brysbaert, M. (2006). Masked phonological priming effects in English: are they real? Do they matter? *Cognition Psychology*, *53*, 97–145.

Rastle, K., McCormick, S. F., Bayliss, L., & Davis, C. J. (2011). Orthography influences the perception and production of speech. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*(*6*), 1588 - 1594.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*, 125–157

Seidenberg, M.S. & Tanenhaus, M.K. (1979). Orthographic effects in rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 546-554.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*, 143–174

Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Mencl, W. E., Fulbright, R. K., Skudlarski, P. … Lyon, G. R., Gore, J. C. (2002). Disruption of posterior brain systems for reading in children with developmental dyslexia. *Biological Psychiatry*, *52*, 101–110.

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

Slobin, D.I. (1973). Cognitive prerequisites for the development of grammar. In C.A. Ferguson & D. I. Slobin (Eds), *Studies of child language development* (pp.175-208). New York: Holt, Rinehart & Winston.

Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry*, 27, 720–31

Snowling, M. (1980). The development of grapheme-phoneme correspondence in normal and dyslexic readers. *Journal of Child Psychology*, *29*, 294-305.

Soares, A. P., Comesaña, M., Iriarte, A., Almeida, J. J., Simões, A., Costa, A., França, P., & Machado, J. (2010). P-PAL: P-PAL: A European Portuguese lexical database, *Linguamática*, *2*(*3*), 67-72.

Soares, A. P., Iriarte, A., Almeida, J. J., Simões, A., Costa, A., França, P., Machado, J., & Comesaña, M. (*in press*). Procura-PALavras (P-PAL): A new measure of word frequency for contemporary European Portuguese. *Psicologia: Reflexão e Crítica*.

Stuart, M., Dixon, M., Masterson, J., & Gray, B. (2003). Children's early reading vocabulary: Description and word frequency lists. *British Journal of Educational Psychology, 73,* 585-598.

Taft, M., Castles, A., Davis, C., Lazendic, G., & Nguyen-Hoan, M. (2008). Automatic activation of orthography in spoken word recognition: Pseudohomograph priming. *Journal of Memory and Language*, *58*, 366–379.

Turkeltaub, P. E., Gareau L., Flowers, D. L., Zeffiro, T.A., & Eden, G. F. (2003). Development of neural mechanisms for reading. *Nature Neuroscience, 6,* 767-773.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.

Unsworth, S. J., & Pexman, P. M. (2003). The impact of reader skill on phonological processing in visual word recognition. *The Quarterly Journal of Experimental Psychology, 56*, 63–81.

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychiatry, 45*(*1*), 2-40.

Ventura P., Morais J., & Kolinsky R. (2006). The development of orthographic consistency effect in speech recognition: From sub-lexical to lexical involvement. *Cognition*, 105, 547–576.

Ventura, P., Kolinsky, R., Pattamadilok, C., & Morais, J. (2008). The developmental turn point of orthographic consistency effects in speech recognition. *Journal of Experimental Child Psychology*, *100*, 135–145.

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *The Canadian Modern Language Review*, 57(*4*), 541-572.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in reading and other tasks. *Journal of Memory and Language*, *47*, 1-29.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*, 3–29.

Ziegler, J. C., & Muneaux, M. (2007). Orthographic facilitation and phonological inhibition in spoken word recognition: A developmental study. *Psychonomic Bulletin & Review*, *14*, 75-80.

Ziegler, J. C., Ferrand, L., & Montant, M. (2004). Visual phonology: The effects of orthographic consistency on different auditory word recognition tasks. *Memory & Cognition*, *32*, 732–741.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, Massachusetts: Addison-Wesley.

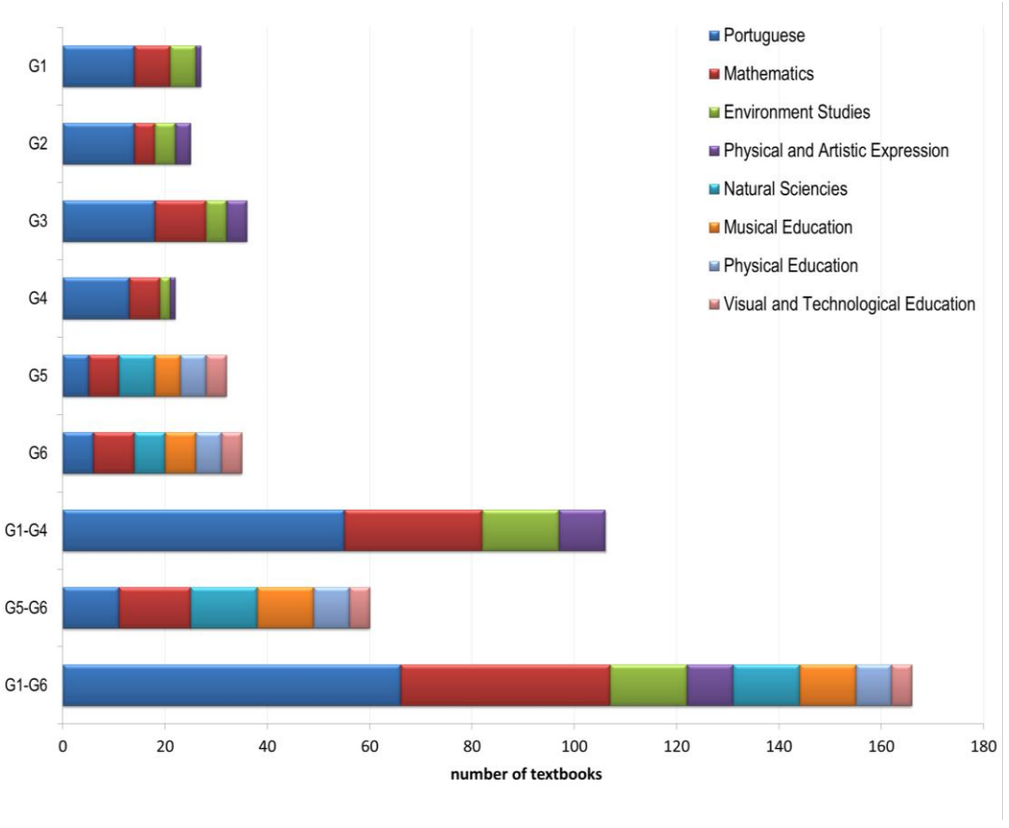

Figure 1. Distribution of the 171 textbooks in ESCOLEX for each grade (G_1 to G_6), educational level (G1-G_4 and (G_5-G_6) and for all school grades combined (G_1-G_6) according to subject area.

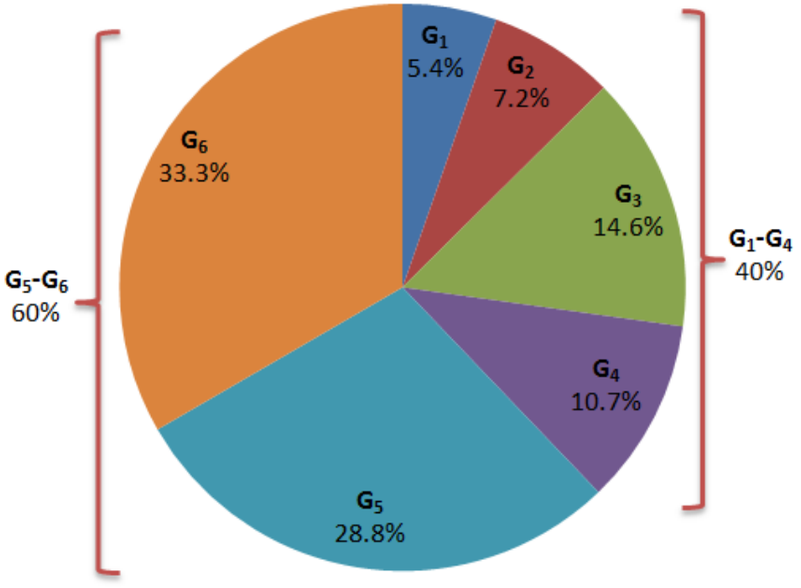

Figure 2. Distribution of the number of occurrences (tokens) in each ESCOLEX subcorpus.

Table 1: Distribution of types, hapax words and words occurring five times or more in each of the nine ESCOLEX grade-levels.

Grade-levels	Word entries (types)	Number hapax words	Number words occurring five or more times
G_1	8,316	2,989	2,894
G_2	13,019	4,558	4,404
G_3	20,444	6,760	7,389
G_4	19,486	7,215	6,267
G_5	31,429	10,149	12,368
G_6	35,216	11,505	13,469
G_1-G_4	29,013	8,817	11,797
G_5-G_6	41,952	12,369	17,577
G_1-G_6	48,381	13,118	22,027

Table 2: Mean, Mode and Percentile values (P_{10}, P_{25}, P_{50}, P_{75} and P_{90}) for word frequency statistics in each ESCOLEX grade-level.

		G_1	G_2	G_3	G_4	G_5	G_6	G_1-G_4	G_5-G_6	G_1-G_6
Overall word frequency (*F*)	Mean	23.08	19.81	25.46	19.33	32.64	33.66	44.04	45.44	65.81
	Mode	1	1	1	1	1	1	1	1	1
	Min	1	1	1	1	1	1	1	1	1
	Max	17,980	15,442	25,055	16,899	45,832	54,252	71,526	86,827	158,353
	P_{10}	1	1	1	1	1	1	1	1	1
	P_{25}	1	1	1	1	1	1	1	1	1
	P_{50}	2	2	3	2	3	3	3	3	4
	P_{75}	8	7	8	7	10	10	11	12	14
	P_{90}	28	24	29	22	37	37	43	47	62
Dispersion across textbooks (*D*)	Mean	.22	.24	.24	,23	.23	.23	.23	.24	.23
	Mode	0	0	0	0	0	0	0	0	0
	Min	0	0	0	0	0	0	0	0	0
	Max	.96	.97	1	1	.96	1	1	1	1
	P_{10}	0	0	0	0	0	0	0	0	0
	P_{25}	0	0	0	0	0	0	0	0	0
	P_{50}	.19	.22	.19	.21	.190	.19	.15	.17	.20
	P_{75}	.40	.41	.41	.43	.41	.39	.39	.41	.39
	P_{90}	.61	.63	.63	.63	.59	.58	.60	.59	.57
Estimated frequency *per*	Mean	89.84	59.25	38.29	39.37	24.08	21.38	27.36	18.36	16.09
	Mode	.28	.28	.10	.15	.06	.08	.01	.03	.01
	Min	.03	.03	.02	.01	.004	.002	.0007	.0009	.0001
	Max	86,285.70	56,938.03	46,464.86	42,597.40	42,284.23	43,178.15	53,555.95	43,122.17	47,049.34
	P_{10}	.15	.15	.06	.13	.03	.02	.009	.01	.003

	P_{25}	.26	.24	.08	.18	.05	.05	.02	.02	.006
	P_{50}	2.53	2.00	.87	1.39	.57	.42	.40	.34	.21
	P_{75}	16.48	11.89	6.39	7.30	4.01	3.26	3.24	2.56	1.74
	P_{90}	84.17	55.10	34.42	37.08	20.20	17.06	19.98	14.06	10.62
Standard frequency index (*SFI*)	Mean	44.04	43.34	40.07	41.65	37.75	37.02	35.62	35.06	32.43
	Mode	34.54	34.54	30.11	31.79	28.04	29.10	21.23	24.97	20.51
	Min	25.17	25.35	21.71	20.66	15.77	13.87	8.70	9.74	.78
	Max	89.36	87.55	86.67	86.29	86.26	86.35	87.29	86.35	86.73
	P_{10}	31.65	31.88	27.79	31.07	24.85	23.71	19.65	20.35	14.07
	P_{25}	34.14	33.87	29.16	32.64	26.66	26.61	21.99	22.48	18.02
	P_{50}	44.03	43.01	39.41	41.42	37.59	36.21	36.06	35.26	33.23
	P_{75}	52.17	50.75	48.05	48.63	46.03	45.13	45.10	44.07	42.41
	P_{90}	59.25	57.41	55.37	55.69	53.05	52.32	53.01	51.48	50.26
Contextual diversity (*CD*)	Mean	.14	.15	.13	.16	.12	.11	.07	.10	.06
	Mode	.04	.04	.03	.05	.03	.03	.01	.02	.01
	Min	.04	.04	.03	.05	.03	.03	.01	.02	.01
	Max	1	1	1	1	1	1	1	1	1
	P_{10}	.04	.04	.03	.05	.03	.03	.01	.02	.01
	P_{25}	.04	.04	.03	.05	.03	.03	.01	.02	.01
	P_{50}	.07	.08	.06	.09	.05	.05	.02	.03	.02
	P_{75}	.15	.16	.14	.18	.14	.12	.07	.09	.05
	P_{90}	.37	.36	.33	.41	.30	.28	.19	.25	.15