LEXIN: A lexical database from Spanish kindergarten and first-grade readers

SILVIA CORRAL, MARTA FERRERO, AND EDURNE GOIKOETXEA
University of Deusto, Bilbao, Spain

The LEXIN database offers psycholinguistic indexes of the 13,184 different words (types) computed from 178,839 occurrences of these words (tokens) contained in a corpus of 134 beginning readers widely used in Spain. This database provides four statistical indicators: *F* (overall word frequency), *D* (index of dispersion across selected readers), *U* (estimated frequency per million words), and *SFI* (standard frequency index). It also gives information about the number of letters, syntactic category, and syllabic structure of the words included. To facilitate comparisons, LEXIN provides data from LEXESP's (Sebastián-Gallés, Martí, Cuetos, & Carreiras, 2000), Alameda and Cuetos's (1995), and Martínez and García's (2004) Spanish adult psycholinguistic frequency databases. Access to the LEXIN database is facilitated by a computer program. The LEXIN program allows for the creation of word lists by letting the user specify searching criteria. LEXIN can be useful for researchers in cognitive psychology, particularly in the areas of psycholinguistics and education.

Psycholinguistic databases collect indexes of psycholinguistic properties of words. This information is very useful for experimental psychologists interested in controlled research with oral and written language stimuli. The use of these psycholinguistic indexes started within the field of applied educational psychology (see, e.g., Thorndike, 1921) and expanded to more basic cognitive research, such as research on visual word recognition.

One important index of printed words is the frequency of occurrence; written frequency has been shown to influence reading accuracy and response times to words greatly. Specifically, research has demonstrated that high-frequency words are responded to more quickly and more accurately than low-frequency words are (Becker, 1976; Forster & Chambers, 1973; for a lexical decision task, see Balota & Chumbley, 1984; for a pioneering work, see Cattell, 1886; for a word-naming task, see Hino & Lupker, 2000; for a review on word frequency effects, see Monsell, 1991). This lexical frequency effect is also evident in the reading performance of neuropsychological patients, such as dyslexics (for a revision, see Behrmann, Plaut, & Nelson, 1998) and patients with Alzheimer's disease (see, e.g., Glosser, Grugan, & Friedman, 1999).

Printed-word frequency counts are usually obtained by collecting words from a representative pool of print resources that are read by a particular group of readers. These counts are presented as frequency dictionaries or frequency norms. Two of the most frequently quoted counts are the Kučera and Francis (1967) dictionary for American English and the CELEX (Baayen, Piepenbrock, & Gulkers, 1995) dictionary for English, Dutch, and German. In Spanish, there are also very widely used counts, such as

the Sebastián-Gallés, Martí, Cuetos, and Carreiras (2000) dictionary (for another dictionary, see also Alameda & Cuetos, 1995; for the addition of other indexes to the pool from Sebastián-Gallés et al., 2000, see Davis & Perea, 2005). All of these counts were obtained from adult print resources.

Similarly, there are many specific counts collected from children's print resources that account for children's reduced visual lexicon (due to their inexperience in reading). The purpose of these children's dictionaries is to offer a more accurate tool for testing initial visual word recognition in this sample. Two of the most frequently quoted counts for primary school grades are *The American Heritage Word Frequency Book* (Carroll, Davies, & Richman, 1971) and *The Educator's Word Frequency Guide* (Zeno, Ivens, Millard, & Duvvuri, 1995), both in American English. Counts in other European languages are the Marconi, Ott, Pesenti, Ratti, and Tavella (1994) dictionary in Italian for primary-school children; the Stuart, Dixon, Masterson, and Gray (2003) dictionary in English for children from 5 to 7 years old; and the Lété, Sprenger-Charolles, and Colé (2004) dictionary in French for primary-school children (for an extension, see Peere-man, Lété, & Sprenger-Charolles, 2007).

In Spanish, there are dictionaries of elementary school children's reading vocabulary (Casanova & Rivera, 1989; Martínez & García, 2004; for an extension, see also Martínez & García, 2008) and productive written vocabulary (Justicia, 1995). However, there is no dictionary for the earliest stages of reading acquisition. This seems surprising if we consider that the vocabulary of the average preschool child ranges between 2,500 and 5,000 lemma or word families (Beck & McKeown, 1991) and that the average

first grader will acquire about 6,000 more (Chall, 1987). And, whereas a preschool reading book includes about 500 word forms, according to our calculations, a first-grade reading book includes about 23,000 words. This accelerated vocabulary growth from preschool to primary school justifies the need for developing a specific dictionary for children who are in the earliest stages of reading acquisition. Thus, the main purpose of this database is to create a specific dictionary for beginning readers.

This dictionary is useful for two purposes. First, using a corpus of words that are common in beginning reading materials makes it possible to create sensitive measures of early reading ability while avoiding the floor effects that usually appear in standardized tests (Bowey, 2005). Second, using this corpus of words provides a more appropriate lexicon for teaching initial word recognition to beginning readers. Therefore, this database is intended to become a useful tool for both researchers and professionals who need an early lexical database of beginning reading materials in Spanish.

Although there is strong evidence for the effects of word frequency, as we stated above, a growing body of literature exists on the issue of whether the effects of word frequency would be better described in terms of the age of acquisition (also termed *order of acquisition*). Age of acquisition, or the age at which words are incorporated into the lexicon, is a variable investigated in accounts of the lexical retrieval and lexical production processes (see, e.g., Barry, Morrison, & Ellis, 1997; Brown & Watson, 1987; Carroll & White, 1973; Gilhooly & Logie, 1980). For example, the effect of age of acquisition on word recognition speed has been demonstrated (for lexical decision tasks, see, e.g., Brysbaert, Lange, & Van Wijnendaele, 2000; Pérez, 2004). This effect has also been shown on other tasks, such as picture naming (e.g., Ellis & Morrison, 1998) and word naming (e.g., Coltheart, Laxon, & Keating, 1988). It seems that the nature of these different tasks influences the size of the age-of-acquisition effect. This effect has been larger in tasks such as picture naming, which involves an arbitrary mapping between picture and phonology, than in tasks such as word naming, which involves a quasiconsistent mapping between orthography and phonology and where it is possible to use what was learned first in later learned words (Lambon-Ralph & Ehsan, 2006; Zevin & Seidenberg, 2002, 2004). Although some authors have argued that it is more parsimonious to explain the effect of the age of acquisition as being one of accumulated frequency, others provide evidence that it is an independent and strong variable (for a revision, see Juhasz, 2005). Because of its controversial status, age of acquisition is a relevant index of printed words for researchers.

Age of acquisition has been estimated using subjective methods, such as adult subjective judgments (e.g., Carroll & White, 1973; Gilhooly & Hay, 1977; Gilhooly & Logie, 1980; Lyons, Teer, & Rubenstein, 1978; Rubin, 1980). However, objective measures, such as objective records of oral production in children, are more accurate (for objective measures in Spanish and English, respectively, see, e.g., Álvarez & Cuetos, 2007, and Morrison, Chappell,

& Ellis, 1997). One objective measure of the age of acquisition of reading vocabulary involves the use of texts specifically designed for readers of certain ages. Thus, a children's frequency dictionary based on kindergarten and first-grade reading material, like the one presented here, can be used not only as a tool for accurately measuring children's early reading vocabulary but also as an objective estimate of the frequency of occurrence of each word in these specific age groups.

In sum, the LEXIN database was created in order to offer a new normative tool for the study of the reading vocabulary of beginning readers in Spanish, a widely used language. In order to facilitate the use of this database, we created software called "LEXIN." This program provides the indexes that have been incorporated into the database and that can be utilized by the user as search criteria for the creation of word lists.

THE LEXIN DATABASE

Corpus Sampling

The LEXIN corpus was compiled from 134 reading and spelling books that were designed for children learning to read (76 for kindergarten and 58 for first grade) by the leading Spanish publishers (see the Appendix for a complete list of readers and additional information). These books are intended to be used in kindergarten or first grade, depending on what each school decides. In Spain, formal reading instruction takes place in the first year of primary school. During the three preschool years (3–6 years old), children are not usually taught how to read. However, preschool literacy instruction varies from school to school. Thus, some teachers try to develop basic decoding skills by introducing some letter–sound correspondences combined with a limited number of sight words.

The reading texts were selected, first, on the basis of sales for the years 2002 and 2003 (per a Santillana representative, personal communication, November 2003). Thus, we computed the cumulative sales figures for the set of readers available for kindergarten and first grade and then retained the sample that accounted for 90% of the sales.

Second, we included readers that used all of the approaches to reading instruction, from code-emphasis approaches to whole-language approaches. However, the majority of the materials are based on a code-emphasis approach, which is the prevailing teaching method in Spain. As Chall (1983) defined it, the code-emphasis approach puts the instructional emphasis on developing learners' recognition of letter–sound correspondences while providing the children with sufficient opportunities to establish their decoding skills. Thus, using a phonics method, the readers present one grapheme at a time and then immediately provide practice in blending the sounds into syllables and whole words. Using a syllabic method, the readers present syllabic families from the beginning (e.g., *ma, me, mi, mo, mu*). Children learn to sound out words by combining known syllables.

Third, we include readers drawn from a broad range of material. We incorporate readers (reading workbooks), narrative texts (stories), and school texts. Therefore, the

sample is reasonably representative of printed Spanish materials for kindergarten and first-grade children.

Frequency Count Computation

We manually copied into a computer all the text data from the readers described above and proofread them. In this process, a word form was considered to be the letters between two blank spaces. Words separated by a dash were included in the database as one entry. The titles of the readers, the names of the authors, and all numerals were omitted.

Next, two computer programs were used. The first one was used to count the frequency of each word in each reader. The second one was used to assign the frequency from the other three Spanish frequency dictionaries to each word and to tag the words for grammatical category. The categories into which the words were classified were taken from the LEXESP dictionary (Sebastián-Gallés et al., 2000): abbreviations, adjectives, adverbs, articles, conjunctions, determiners, interjections, nouns, numerals (cardinal and ordinal numerals, not Arabic numerals), prepositions, pronouns, residuals, and verbs.

Finally, two people carried out an extensive editing of the input files manually. Any printing or spelling errors in the words were corrected using the electronic consultation system of the Web page of the Real Academia Española (RAE, 2001). To facilitate the recording (computation) of those words that involved some difficulty, the following criteria were adopted.

Capitals. We converted uppercase letters to lowercase, even in the case of proper nouns.

Individual letters and syllables. We included the names of all the letters in the alphabet, both vowels and consonants, except *w*, since its name is a compound word. However, we eliminated the nonsense syllables designed to show letter associations (e.g., *ba*, *dro*).

Slang. We respected the slang that appeared in the children's books, such as "enfadao" (*enfadado*), as well as the shortened forms, such as "profe" (*profesor*) and "tele" (*televisión*).

Diminutives and augmentatives. On the basis of the suffixes and prefixes included in the RAE dictionary, we included all those words that indicate diminutive and augmentative conditions (e.g., *quesito*, *gatazo*, *supermercado*).

Invented or fictional feminine or masculine words. Words were included that were formed according to the usual rules for forming the masculine and feminine, even when they did not have an entry in the RAE dictionary (e.g., *azafato*).

Prefixes and suffixes. All the prefixes and suffixes that accompany a word (e.g., *desilusión*, *cucharada*) were included, and those that appeared alone (*pre-*, *-dad*) were excluded. All the words that were properly formed by a prefix or suffix were included, even if they were not known words (*superpollo*).

Foreign words. We respected the words coming from other languages, regardless of whether they had an entry or a hispanicized version in the RAE dictionary (e.g., *walkman*, *anorak*).

Finally, although the majority of the words had an entry in the RAE dictionary, some nonsense words were also included, as were misspellings that did not meet this condition. These are not the results of typing or scanning errors in the input; those were carefully edited out. These nonwords and misspellings resulted instead from several other sources, such as improperly written onomatopoeias (e.g., *toc*, instead of *tac*), expressions from the children's colloquial language itself (e.g., *michín*, *yupi*), and incorrectly written interjections (i.e., *hmm*, instead of *hum*). In the latter case, the errors were corrected if they consisted of repetitions and not of substitutions (e.g., *ayyy*, instead of *ay*). We did not eliminate these nonsense words and misspellings from the database because, although they have no meaning, they form a part of the written material directed toward children.

Description of the File

The LEXIN database and the associated program can be downloaded by anonymous file transfer (<http://paginaspersonales.deusto.es/egoiko/>).

Our database includes 13,184 words. Each word is followed by 10 columns corresponding to the 10 indexes described below. The frequency indexes were computed following the methods first described by Carroll et al. (1971) and more recently by Breland (1996) and Lété et al. (2004).

Frequency (F) is the number of occurrences of each word. *Dispersion (D)* is the dispersion or distribution of the frequency of each word, across readers. *D* ranges from .00 to 1.00 and is equal to .00 when all occurrences of the word are found in a single reader, regardless of the frequency. *D* is equal to 1.00 if the frequencies are distributed in exactly equal proportions across readers. The formula for calculating *D* is

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

where *n* is the number of readers in the corpus, *i* is the reader number (1, 2, . . . , *n*), and *p_i* is the frequency of a word in the *i*th reader, with *p_i log p_i* = 0, if *p_i* = 0.

Frequency per million (U) is the estimated frequency of each word per million words adjusted for *D*. When *D* = 1, *U* is computed simply as the frequency per million words. But when *D* < 1, the value of *U* is adjusted downward. When *D* = 0, *U* has a minimum value that is based on the average weighted probability of the word's occurrence across all of the readers. 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 the corpus (13,184), *F* is the frequency of the word in the corpus, *D* is the index of dispersion, and *f_{min}* is 1/*N* times the sum of the products of *f_i* and *s_i*, *f_i* is the frequency in Reader *i*, and *s_i* is the number of words in that reader.

Standard frequency index (SFI) is derived directly from *U*. As Lété et al. (2004) pointed out, the user should find this index to be a simple and convenient way of indicating frequency counts. Thus, for example, a value that can serve as a reference when using this index is the *SFI* of 40,

which corresponds to the value for a word that occurs once in a million words. Other values that can be of practical use are an *SFI* of 70, which corresponds to words that can be expected to occur once in every 1,000 words, and an *SFI* of 90, which corresponds to words that are expected to occur once in every 10 words, and so forth. The *SFI* is computed from *U* by using the formula

$$SFI = 10 * [\log_{10}(U) + 4].$$

Taking, for example, the words *leer* and *bayas*, it is possible to see the use of the indexes described above. Both of the words have the same frequency (48), but they have different *D* values (.57 and .04, respectively). Their respective estimated frequencies per million are 13,152 and 2,262. Consequently, the *SFI* values are 81.19 and 73.54, respectively.

N letters is the number of letters in each word. *Structure* is the syllabic structure of each word. We used a syllabication algorithm that followed the rules for syllabifying Spanish language words, as stated by the RAE (1999).

LEXESP is the frequency of each word in the LEXESP database (Sebastián-Gallés et al., 2000). LEXESP is a frequency database that is based on a count of approximately 5 million Spanish words and includes indexes such as number of syllables, stress location, pronunciation, imageability, concreteness, and familiarity, among others.

A&C is the frequency of each word in Alameda and Cuetos's (1995) database. Alameda and Cuetos's frequency dictionary is based on a count of approximately 2 million Spanish words.

M&G is the frequency of each word in Martínez and García's (2004) database. Martínez and García's frequency dictionary has a total of approximately 100,000 words selected from the books that a small group of children from 6 to 12 years of age read during the year.

Category refers to the syntactic categories that have been included in the database—namely, abbreviations, adjectives, adverbs, articles, conjunctions, determiners, interjections, nouns, numerals (cardinal and ordinal numerals, not Arabic numerals), prepositions, pronouns, residuals, and verbs. In turn, we included syntactic subcategories whose attributes varied according to the syntactic category to which they referred. Both the categories and the subcategories were taken from the LEXESP database.

Grade refers to the school grade (kindergarten or first grade) in which it is probable that children will encounter the word for the first time, according to the publishers' suggested use for the readers. Note that it is only a recommended use and that, in fact, several readers are intended for use by both kindergartners and first graders.

Description of the Program

The LEXIN program was written in Java. It is a multiplatform program with a .jar file for every platform and an .exe file for the Windows platform. It is menu driven for all options, which makes it easy for novices to use. Help is available in a separate Word file. The Help file covers the running of the program.

When starting to use the database, the user has to choose from a menu containing two options. (1) The *archive* option

contains the different operations the user can perform with the application: *new list* (to include a new list of words in the database that can later be consulted), *search* (to perform searches using different criteria from the Word lists already included in the database), and *exit* (to stop running the program). (2) By means of the *language* option, the user can select the language with which to work in the application. The user can choose from among Spanish, English, or Basque. The separate Help Word file contains the necessary instructions for exploring these options.

Hardware Specifications

The program will run on any IBM-compatible (Pentium) computer with any operating system (e.g., Windows, Red Hat, OS/2). The program itself amounts to approximately 3.3 MB, and the LEXIN database amounts to approximately 6.4 MB.

Descriptive Statistics

LEXIN contains 93,514 characters divided into 13,184 different words (types) and 178,839 occurrences of these words (tokens). From the total, 6,912 words are included in readers recommended for kindergarten (but not exclusively kindergarten) children, and 6,276 words are included in readers for first graders. Like other databases in other languages with counts based on word forms (e.g., Carroll et al., 1971; Stuart et al., 2003), the most obvious characteristic is the bias toward the lower frequencies. Thus, the 100 most frequently occurring words (less than 1%) account for 44.53% of all tokens. The 500 most frequently occurring words (3.79%) account for 61.77% of all tokens. The fact that such a reduced number of words makes up a large part of the total frequency shows an irregular distribution of frequencies in the set of words. Furthermore, the proportion of hapax (one occurrence) words in this database—almost a third of the words (3,644 words, 27.6%)—is also a reflection of this lack of balance. This proportion is not as large as that of the majority of the corpora, where hapax words usually represent 50% of the total corpus (e.g., Carroll et al., 1971; Stuart et al., 2003). Moreover, 71% of the words have low frequencies (below 5 occurrences). This high percentage of low-frequency words in readers for beginners can represent a problem that was previously pointed out by Stuart et al.—that is, that children do not see some words repeated enough times to be able to learn them.

The most frequent grammatical categories in LEXIN are nouns (46.09%), verbs (33.06%), and adjectives (18.39%); the less frequent grammatical categories are interjections (0.02%), pronouns (0.15%), and conjunctions (0.2%). However, when lexical frequency is taken into account, our database confirms that the most frequent words in the early reading vocabulary are the function words (i.e., articles, prepositions, adverbs, pronouns), rather than the content words, just as occurs in children's early reading vocabulary in English (Stuart et al., 2003). Table 1 shows the syntactic category of the 100 most frequent words in the database, of which function words alone account for 39.11%. Nevertheless, just as in the English database (Stuart et al., 2003), as the frequency token decreases, there is an increase in the percentage of content

Table 1
Syntactic Categories of the 100 Most Frequent Words in the LEXIN Database

Syntactic Category	N	Items	Classification
Definite article	5	la, el, los, las, lo	Function
Indefinite article	2	un, una	Function
Conjunctions	5	y, cuando, pero, o, porque	Function
Determiners	—	—	Function
Prepositions	9	de, a, en, para, por, sin, si, hasta, con	Function
Pronouns	12	que, se, le, qué, me, yo, te, nos, ese, les, esta, quién	Function
Adverbs	10	no, muy, más, como, ya, cómo, sí, así, también, después	Function
Verbs	23	es, tiene, está, escribe, ha, son, hay, lee, era, soy, hace, dijo, va, había, estaba, da, dice, rodea, tengo, colorea, lleva, hacer, están	Function
Contractions	2	al, del	Function
Interjections	—	—	Function
Nouns	20	casa, mamá, día, papá, sol, luna, gato, agua, bien, monstruo, abuelo, niño, palabra, niños, nombre, perro, mar, niña, animales, mesa	Content
Proper noun	1	Ana	Content
Adjectives	11	su, mi, sus, dos, todos, completa, todo, mucho, tu, cada, tres	Content

Note—Many words in this table have multiple classifications, with the possibility of being classified into two or more categories (e.g., “bien” can be a noun, conjunction, or adverb). Here, we chose the first classification given by RAE.

words, until they come to represent 99% of the last 100 words of the first thousand.

With regard to the length of the words in the beginning reading vocabulary, approximately 75% of the words have between 5 and 9 letters, with a mean length of 7.09 letters per word ($SD = 2.16$). Less than 10% of the words have 4 letters or less, and only 2.7% have 3 letters or less, unlike other languages, like English, with shorter words and a greater number of monosyllables (Fenk-Oczlon & Fenk, 2008).

Another characteristic of the Spanish beginning vocabulary is the variety of syllabic structures. Evidence of this lies in the seven different syllabic structures observed in the 100 most frequent monosyllables. These are listed in order of frequency in Table 2. As Table 2 also shows, the most common structures are the CVC and CV syllables, as is the case in English (Stuart et al., 2003), followed by CVV and then VC structures.

Conclusion

There are no previous word dictionaries that have measured the frequency of words in beginning reader materials in Spanish. However, the importance of the printed-

word frequency count in beginning readers is clear for several reasons. Frequency dictionaries for children will make it possible to control the frequency of words, which, as stated earlier, is one of the most influential characteristics on several tasks related to visual and oral language processing research. Consequently, this information is very useful in developmental reading research. Another important use of dictionaries for children is in the fields of applied and educational assessment and teaching. Children's frequency dictionaries can guide the selection and sequencing of target language features for language assessment and teaching/instruction. A third important use is the possible application of this database for adult reading research that aims to employ printed words with a very early age of acquisition. Thus, the LEXIN database can be a useful tool for linguistic and psycholinguistic research, as well as for teachers and other education professionals.

The development of a software program to expedite the creation of word lists discussed in this study would have a huge beneficial impact on the use of LEXIN, both in research and in professional settings.

One limitation of the present database is that the usefulness of the results and their possible generalization might

Table 2
Syllabic Structure of the 100 Most Frequent Monosyllables, in Descending Order of Frequency of the Structure

Structure	Items	Total
CVC	los, las, con, del, por, muy, sus, más, tos, son, hay, nos, sol, soy, sin, les, mar, tan, ves, hoy, ver, han, ser, rey, van, pan, mis, voy, luz, sal, has, mal, tor, dar, pez, ven, don, tus, fin, vos	40
CV	la, de, se, no, su, le, me, lo, mi, yo, ha, te, si, ya, va, sí, tu, da, tú, ni, he, mí, ve, sé, ti, mu, ja	27
CVV	que, qué, lee, día, fue, río, veo, dio, pie, pío, tía, tío, vio, feo, zoo, ría	16
VC	el, en, un, es, al, él, ir, ay, os	9
CCVC	tres, flor, tren, gran, flan	5
V	a, o	2
C	y	1

be limited to the particular nature of the sample—namely, Spanish-speaking children from Spain. The question of whether similar results would be obtained from a sample of Spanish-speaking children outside of Spain requires further research. Another issue is that it may not be an objective measure of the age of acquisition, due to a possible cohort effect, since the words were obtained from a specific sample and may not correspond to the age of acquisition of previous or future generations. However, some studies indicate that a cohort effect occurs only with those words that fall out of use, refer to technological advances, or stem from a different lifestyle (see Bird, Franklin, & Howard, 2001).

In the future, the task of maintaining the database will be necessary, as will the analysis of the quantity and quality of the vocabulary directed toward children in textbooks. Another issue of interest for the future is the exploration of other lexical and infralexical variables not included here—particularly, the index of grapheme–phoneme and phoneme–grapheme consistency, which differs in several languages, thus affecting literacy acquisition.

AUTHOR NOTE

This research was supported in part by Grant HU2006-13 from the Departamento de Educación, Universidades e Investigación del Gobierno Vasco. We are grateful to the authors of the cited Spanish databases for allowing us to draw values for the present words, and to Bernard Lété for providing us with invaluable material. We also thank Cindy De Poy and Mari Luz Guenaga for helping us with the English and electronic languages, respectively. Correspondence concerning this article should be addressed to E. Goikoetxea, Departamento de Psicopedagogía, Universidad de Deusto, Apartado 1, 48080-Bilbao, Spain (e-mail: edurne.goikoetxea@deusto.es).

REFERENCES

- ALAMEDA, J. R., & CUETOS, F. (1995). *Diccionario de frecuencias de las unidades lingüísticas del castellano* [Frequency dictionary of Spanish linguistic units]. Oviedo, Spain: Servicio de Publicaciones de la Universidad de Oviedo.
- ÁLVAREZ, B., & CUETOS, F. (2007). Objective age of acquisition norms for a set of 328 words in Spanish. *Behavior Research Methods*, *39*, 377-383.
- BAAYEN, R. H., PIEPENBROCK, R., & GULIKERS, L. (1995). *The CELEX lexical database* (Release 2) [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.
- BALOTA, D. A., & CHUMBLEY, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance*, *10*, 340-357. doi:10.1037/h0084192
- BARRY, C., MORRISON, C. M., & ELLIS, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency and name agreement. *Quarterly Journal of Experimental Psychology*, *50A*, 560-585. doi:10.1080/027249897392026
- BECK, I. L., & MCKEOWN, M. G. (1991). Social studies texts are hard to understand: Mediating some of the difficulties. *Language Arts*, *68*, 482-490.
- BECKER, C. A. (1976). Allocation of attention during visual word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, *2*, 556-566. doi:10.1037/0096-1523.2.4.556
- BEHRMANN, M., PLAUT, D. C., & NELSON, J. (1998). A literature review and new data supporting an interactive account of letter-by-letter reading. *Cognitive Neuropsychology*, *15*, 7-51. doi:10.1080/026432998381212
- BIRD, H., FRANKLIN, S., & HOWARD, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers*, *33*, 73-79.
- BOWEY, J. A. (2005). Predicting individual differences in learning to read. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 155-171). Malden, MA: Blackwell.
- BRELAND, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science*, *7*, 96-99. doi:10.1111/j.1467-9280.1996.tb00336.x
- BROWN, G. D. A., & WATSON, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition*, *15*, 208-216.
- BRYLSBAERT, M., LANGE, M., & VAN WIJNENDEAELE, I. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, *12*, 65-85. doi:10.1080/095414400382208
- CARROLL, J. B., DAVIES, P., & RICHMAN, B. (Eds.) (1971). *The American Heritage word frequency book*. Boston: Houghton Mifflin.
- CARROLL, J. B., & WHITE, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, *25*, 85-95.
- CASANOVA, M. A., & RIVERA, M. (1989). *Vocabulario básico en la E.G.B.* [Basic vocabulary in primary school]. Madrid: Ministerio de Educación y Ciencia.
- CATTELL, J. M. (1886). The time taken up by cerebral operations. *Mind*, *11*, 220-242, 377-392, 524-538.
- CHALL, J. S. (1983). *Learning to read: The great debate* (2nd ed.). New York: Harcourt Brace.
- CHALL, J. S. (1987). Two vocabularies for reading: Recognition and meaning. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 7-17). Hillsdale, NJ: Erlbaum.
- COLTHEART, V., LAXON, V. J., & KEATING, C. (1988). Effects of word imageability and age of acquisition on children's reading. *British Journal of Psychology*, *79*, 1-12.
- DAVIS, C. J., & PEREA, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, *37*, 665-671.
- ELLIS, A. W., & MORRISON, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 515-523. doi:10.1037/0278-7393.24.2.515
- 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.
- FORSTER, K. I., & CHAMBERS, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning & Verbal Behavior*, *12*, 627-635. doi:10.1016/S0022-5371(73)80042-8
- GILHOOLY, K. J., & HAY, D. (1977). Imagery, concreteness, age-of-acquisition, familiarity, and meaningfulness values for 205 five-letter words having single-solution anagrams. *Behavior Research Methods & Instrumentation*, *9*, 12-17.
- GILHOOLY, K. J., & LOGIE, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12*, 395-427.
- GLOSSER, G., GRUGAN, P., & FRIEDMAN, R. B. (1999). Comparison of reading and spelling in patients with probable Alzheimer's disease. *Neuropsychology*, *13*, 350-358. doi:10.1037/0894-4105.13.3.350
- HINO, Y., & LUPKER, S. J. (2000). The effects of word frequency and spelling-to-sound regularity in naming with and without preceding lexical decision. *Journal of Experimental Psychology: Human Perception & Performance*, *26*, 166-183. doi:10.1037/0096-1523.26.1.166
- JUHÁSZ, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, *131*, 684-712. doi:10.1037/0033-2909.131.5.684
- JUSTICIA, F. (1995). *El desarrollo del vocabulario: Diccionario de frecuencias* [Developmental vocabulary: Frequency dictionary]. Granada: Universidad de Granada.
- KUCERA, H., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- LAMBON-RALPH, M. A., & EHSAN, S. (2006). Age of acquisition effects depend on the mapping between representations and the frequency of occurrence: Empirical and computational evidence. *Visual Cognition*, *13*, 928-948. doi:10.1080/13506280544000110
- LÉTÉ, B., SPRENGER-CHAROLLES, L., & COLÉ, P. (2004). MANULEX:

- A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, **36**, 156-166.
- LYONS, A. W., TEER, P., & RUBENSTEIN, H. (1978). Age-at-acquisition and word recognition. *Journal of Psycholinguistic Research*, **7**, 179-187. doi:10.1007/BF01067041
- MARCONI, L., OTT, M., PESENTI, E., RATTI, D., & TAVELLA, M. (1994). *Lessico elementare: Dati statistici sull'italiano scritto e letto dai bambini delle elementari* [Elementary lexicon: Statistical data for Italian written and read by elementary school children]. Bologna: Zanichelli.
- MARTÍNEZ, J. A., & GARCÍA, M. E. (2004). *Diccionario de frecuencias del castellano escrito en niños de 6 a 12 años* [Dictionary of frequencies of written Spanish in 6- to 12-year-old children]. Salamanca: Universidad Pontificia de Salamanca.
- MARTÍNEZ, J. A., & GARCÍA, M. E. (2008). ONESC: A database of orthographic neighbors for Spanish read by children. *Behavior Research Methods*, **40**, 191-197.
- 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.
- MORRISON, C. M., CHAPPELL, T. D., & ELLIS, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal of Experimental Psychology*, **50A**, 528-559. doi:10.1080/02729897392017
- PEEREMAN, R., LÉTÉ, B., & SPRENGER-CHAROLLES, L. (2007). Manulex-infra: Distributional characteristics of grapheme-phoneme mappings, and infralexical and lexical units in child-directed written material. *Behavior Research Methods*, **39**, 579-589.
- PÉREZ, M. A. (2004). Influencia del orden de adquisición del léxico en el reconocimiento de palabras [Influence of lexical order-of-acquisition on word recognition]. Unpublished doctoral dissertation, Universidad de Murcia.
- REAL ACADEMIA ESPAÑOLA (1999). *Ortografía de la lengua española* [Spanish language orthography]. Madrid: Espasa Calpe.
- REAL ACADEMIA ESPAÑOLA (2001). *Diccionario de la lengua española* [Spanish language dictionary]. Retrieved from www.rae.es/rae.html.
- RUBIN, D. C. (1980). 51 properties of 125 words: A unit analysis of verbal behavior. *Journal of Verbal Learning & Verbal Behavior*, **19**, 736-755. doi:10.1016/S0022-5371(80)90415-6
- SEBASTIÁN-GALLÉS, N., MARTÍ, M. A., CUETOS, F., & CARREIRAS, M. F. (2000). *LEXESP: Léxico informatizado del español* [LEXESP: A computerized word-pool in Spanish]. Barcelona: Edicions de la Universitat de Barcelona.
- 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. doi:10.1348/000709903322591253
- THORNDIKE, E. L. (1921). *The teacher's word book*. New York: Columbia University, Teachers College Press.
- ZENO, S. M., IVENS, S. H., MILLARD, R. T., & DUVVURI, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.
- ZEVIN, J. D., & SEIDENBERG, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory & Language*, **47**, 1-29. doi:10.1006/jmla.2001.2834
- ZEVIN, J. D., & SEIDENBERG, M. S. (2004). Age of acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory & Cognition*, **32**, 31-38.

APPENDIX

List of the Readers in the LEXIN Corpus

	Title	Grade	Publisher
1.	Lectoescritura 1	Kindergarten	Algaida
2.	Lectoescritura 2	Kindergarten	Algaida
3.	Lectoescritura 3	Kindergarten	Algaida
4.	Lectoescritura 4	Kindergarten	Algaida
5.	Lectoescritura 5	Kindergarten	Algaida
6.	Lectoescritura. Pecosete y Pecoseta. Consonantes 1	Kindergarten	Algaida
7.	Lectoescritura. Pecosete y Pecoseta. Consonantes 2	Kindergarten	Algaida
8.	Lectoescritura. Pecosete y Pecoseta. Consonantes 3	Kindergarten	Algaida
9.	Lectoescritura. Pecosete y Pecoseta. Consonantes 4	Kindergarten	Algaida
10.	1ª Cartilla. Nuevo Palau	Kindergarten	Anaya
11.	2ª Cartilla. Nuevo Palau	Kindergarten	Anaya
12.	3ª Cartilla. Nuevo Palau	Kindergarten	Anaya
13.	Proyecto siete colores: Aprendo a leer	Kindergarten	Anaya
14.	Poquito a poco	Kindergarten	Anaya
15.	Empiezo a leer	Kindergarten	Anaya
16.	Toc, Toc, ábreme. Lecturas	First grade	Anaya
17.	Lecturas 1. Ventana de colores	First grade	Anaya
18.	Poquito a poco. Cuaderno 1	First grade	Anaya
19.	Poquito a poco. Cuaderno 2	First grade	Anaya
20.	Poquito a poco. Cuaderno 3	First grade	Anaya
21.	Micho 1	Kindergarten	Bruño
22.	Micho 2	Kindergarten	Bruño
23.	Colección BEABÁ/1. 1	Kindergarten	Casals
24.	Colección BEABÁ/1. 2	Kindergarten	Casals
25.	Colección BEABÁ/1. 3	Kindergarten	Casals
26.	Colección BEABÁ/1. 4	Kindergarten	Casals
27.	Colección BEABÁ/1. 5	Kindergarten	Casals
28.	Colección BEABÁ/2. 1	Kindergarten	Casals
29.	Colección BEABÁ/2. 2	Kindergarten	Casals
30.	Colección BEABÁ/2. 3	Kindergarten	Casals
31.	Colección BEABÁ/2. 4	Kindergarten	Casals

APPENDIX (Continued)

	Title	Grade	Publisher
32.	Colección BEABÁ/2. 5	Kindergarten	Casals
33.	Colección BEABÁ/2. 6	Kindergarten	Casals
34.	Colección BEABÁ/2. 7	Kindergarten	Casals
35.	Colección BEABÁ/2. 8	Kindergarten	Casals
36.	Colección BEABÁ/2. 9	Kindergarten	Casals
37.	Colección BEABÁ/2. Lecturas. Tor y Tuga	Kindergarten	Casals
38.	Casals. Libro de lengua. 1º	First grade	Casals
39.	Casals. Cuaderno de actividades 0. 1º	First grade	Casals
40.	Casals. Cuaderno de actividades 1. 1º	First grade	Casals
41.	Casals. Cuaderno de actividades 2. 1º	First grade	Casals
42.	Casals. Cuaderno de actividades 3. 1º	First grade	Casals
43.	Casals. Cuaderno de enlace 1. 1º	First grade	Casals
44.	Casals. Cuaderno de enlace 2. 1º	First grade	Casals
45.	Proyecto Cosquillas. Lectoescritura. Cuadernos 1–10	Kindergarten	Edebé
46.	Lengua y Literatura 1	First grade	Edebé
47.	Cuaderno de Lengua 1	First grade	Edebé
48.	Cuaderno de Lengua 2	First grade	Edebé
49.	Cuaderno de Lengua 3	First grade	Edebé
50.	Érase una vez el país de las letras 1	Kindergarten	Edelvives
51.	Érase una vez el país de las letras 2	Kindergarten	Edelvives
52.	Érase una vez el país de las letras 3	Kindergarten	Edelvives
53.	Érase una vez el país de las letras 4	Kindergarten	Edelvives
54.	Lengua 1º	First grade	Edelvives
55.	Imaginario Lecturas 1º	First grade	Edelvives
56.	Cuaderno de actividades 1º	First grade	Edelvives
57.	Proyecto Ágora. Lengua	First grade	Everest
58.	Proyecto Ágora. Cuadernillo de evaluación	First grade	Everest
59.	Proyecto Ágora. Cuadernillo de refuerzo y ampliación	First grade	Everest
60.	Proyecto Luna. Cuadernillo 1. Primer trimestre	Kindergarten	Everest
61.	Proyecto Luna. Cuadernillo 2. Segundo trimestre	Kindergarten	Everest
62.	Proyecto Luna. Cuadernillo 3. Tercer trimestre	Kindergarten	Everest
63.	Proyecto Fantasía. Cuadernillo 1. Primer trimestre	Kindergarten	Everest
64.	Proyecto Fantasía. Cuadernillo 2. Segundo trimestre	Kindergarten	Everest
65.	Proyecto Fantasía. Cuadernillo 3. Tercer trimestre	Kindergarten	Everest
66.	Proyecto Nuevo Flopi. Cuadernillo 1. Primer trimestre	Kindergarten	Everest
67.	Proyecto Nuevo Flopi. Cuadernillo 2. Segundo trimestre	Kindergarten	Everest
68.	Proyecto Nuevo Flopi. Cuadernillo 3. Tercer trimestre	Kindergarten	Everest
69.	1ª Cartilla	Kindergarten	Lamela
70.	2ª Cartilla	Kindergarten	Lamela
71.	3ª Cartilla	Kindergarten	Lamela
72.	Escritura cuaderno 1. Nivel 1	Kindergarten	Santillana
73.	Escritura cuaderno 2. Nivel 1	Kindergarten	Santillana
74.	Escritura cuaderno 1. Nivel 2	Kindergarten	Santillana
75.	Escritura cuaderno 2. Nivel 2	Kindergarten	Santillana
76.	Lectura 1	Kindergarten	Santillana
77.	Lectura 2	Kindergarten	Santillana
78.	Chinchirimbola nº 1	Kindergarten	Santillana
79.	Chinchirimbola nº 2	Kindergarten	Santillana
80.	Chinchirimbola nº 3	Kindergarten	Santillana
81.	Chinchirimbola nº 4	Kindergarten	Santillana
82.	Cuentos de la luna lunera	First grade	Santillana
83.	Luna Lunera. Nº 1	First grade	Santillana
84.	Luna Lunera. Nº 2	First grade	Santillana
85.	Luna Lunera. Nº 3	First grade	Santillana
86.	Luna Lunera. Nº 4	First grade	Santillana
87.	Luna Lunera. Nº 5	First grade	Santillana
88.	Luna Lunera. Nº 6	First grade	Santillana
89.	Luna Lunera. Nº 7	First grade	Santillana
90.	Luna Lunera. Nº 8	First grade	Santillana
91.	Luna Lunera. Nº 9	First grade	Santillana
92.	Luna Lunera. Nº 10	First grade	Santillana
93.	Luna Lunera. Cuaderno de escritura nivel 1	First grade	Santillana
94.	Luna Lunera. Cuaderno de escritura nivel 2	First grade	Santillana

Appendix (Continued)

Title	Grade	Publisher
95. Lengua Castellana. Nuestro mundo. La granja	First grade	Santillana
96. Lengua Castellana. El bosque de los cuentos	First grade	Santillana
97. Cuaderno de lengua castellana. Fichas de lectura. 1 ^{er} trimestre	First grade	Santillana
98. Cuaderno de lengua castellana. Fichas de lectura. 2 ^o trimestre	First grade	Santillana
99. Cuaderno de lengua castellana. Fichas de lectura. 3 ^{er} trimestre	First grade	Santillana
100. Cuaderno de lengua castellana. Fichas de escritura. 1 ^{er} trimestre	First grade	Santillana
101. Cuaderno de lengua castellana. Fichas de escritura. 2 ^o trimestre	First grade	Santillana
102. Cuaderno de lengua castellana. Fichas de escritura. 3 ^{er} trimestre	First grade	Santillana
103. Lecturas amigas. En marcha	First grade	Santillana
104. Lecturas amigas. Primeros pasos	First grade	Santillana
105. La cartilla	First grade	Santillana
106. Letras encantadas. Lectoescritura	Kindergarten	Santillana
107. Letras encantadas. Lectoescritura 1	Kindergarten	Santillana
108. Letras encantadas. Lectoescritura 2	Kindergarten	Santillana
109. Letras encantadas. Lectoescritura 3	Kindergarten	Santillana
110. Letras encantadas. Lectoescritura 4	Kindergarten	Santillana
111. Letras encantadas. Lectoescritura 5	Kindergarten	Santillana
112. Letras encantadas. Lectoescritura 6	Kindergarten	Santillana
113. Ven a leer 1	Kindergarten	Siglo XXI
114. Ven a leer 2	Kindergarten	Siglo XXI
115. Ven a leer 3	Kindergarten	Siglo XXI
116. Iniciación a la lectura 1	Kindergarten	SM
117. Iniciación a la lectura 2	Kindergarten	SM
118. Iniciación a la escritura 1	Kindergarten	SM
119. Iniciación a la escritura 2	Kindergarten	SM
120. Proyecto Duendes. Lecturas 1	First grade	SM
121. Proyecto Duendes. 1 ^{er} Trimestre	First grade	SM
122. Proyecto Duendes. 2 ^o Trimestre	First grade	SM
123. Proyecto Duendes. 3 ^{er} Trimestre	First grade	SM
124. Proyecto Duendes. El álbum de las palabras	First grade	SM
125. Proyecto Duendes. Cuaderno de lengua 1 ^{er} trimestre	First grade	SM
126. Proyecto Duendes. Cuaderno de lengua 2 ^o trimestre	First grade	SM
127. Proyecto Duendes. Cuaderno de lengua 3 ^{er} trimestre	First grade	SM
128. Proyecto Papelo. Lengua 1 ^{er} curso	First grade	SM
129. Proyecto Papelo. Escribir 1	First grade	SM
130. Proyecto Papelo. Escribir 2	First grade	SM
131. Proyecto Papelo. Lecturas	First grade	SM
132. Vamos a jugar 1	Kindergarten	Vicens vives
133. Leo y escribo	First grade	Vicens vives
134. Vamos a leer 1	Kindergarten	Vicens vives

(Manuscript received December 10, 2008;
revision accepted for publication May 14, 2009.)