CoDAS
ISSN 2317-1782 (Online version)

Original Article
Artigo Original

Words database for reading by students from Basic Education I, E-READING I

Banco de palavras para leitura de escolares do Ensino Fundamental ciclo I, E-LEITURA I

Adriana Marques de Oliveira[1]
Jair Licio Ferreira Santos[2]
Simone Aparecida Capellini[1]

Keywords

Reading
Assessment
Education, Primary
Learning
Education Status

ABSTRACT

Purpose: To present the process of elaborating a words database appropriate for the reading proficiency level of elementary school students. **Methods:** Words from Portuguese language textbooks used in the public school system of São Paulo, Paraná, Rio de Janeiro and Minas Gerais states of Brazil were selected. We opted for those belonging to the class of nouns and adjectives. Were excluded: homophones; other languages; abbreviations; adverbs; adverbial phrases; prepositional phrases; months of the year; numerals; diminutive or augmentative forms; proper names; misspellings; slang; and words composed by juxtaposition. The words were then categorized according to frequency of occurrence in the textbooks. For this purpose, the tertiles of the distribution, the mean frequency and cutoff point of the tertiles were used. To detect possible mistakes in the words selection, 50 students from the 1st to 5th year, 10 per school year, were selected for individual reading from the database for 20 minutes. **Results:** A total of 286,290 words were typed. After analyzing the inclusion/exclusion criteria and categorizing by frequency of occurrence, the database amounted to 4,195 words. Following the students reading, the E-READING I comprised 4,190 words classified according to frequency: low (n = 3735), medium (n = 374) and high (n = 81). **Conclusion:** The development of a low, medium and high frequency words database, to serve as a linguistic stimulus, was achieved and made available for clinical and pedagogical practice.

Descritores

Leitura
Avaliação
Ensino Fundamental
Aprendizagem
Escolaridade

RESUMO

Objetivo: Apresentar o processo de elaboração de um banco de palavras adequadas ao nível de proficiência de leitura de escolares do Ensino Fundamental I. **Método:** Selecionaram-se palavras de livros didáticos de Língua Portuguesa da rede pública de ensino de São Paulo, Paraná, Rio de Janeiro e Minas Gerais. Optou-se pelos substantivos e adjetivos. Excluíram-se as palavras homófonas, escritas em outros idiomas, com grafia errada, compostas por justaposição, abreviações, advérbios, locuções adverbiais, locuções prepositivas, meses do ano, numerais, palavras no aumentativo ou diminutivo, nomes próprios e gírias. As palavras foram categorizadas segundo frequência de ocorrência nos livros. Para tanto, foram utilizados os tercis da distribuição, a frequência média e o ponto de corte dos tercis. Para detectar possíveis falhas na seleção das palavras, foram selecionados 50 escolares do 1º ao 5º ano (10 por ano escolar) para leitura individual, com duração de 20 minutos, do banco de palavras. **Resultado:** Foram digitadas 286.290 palavras. Após análise dos critérios de inclusão/exclusão e categorização por frequência de ocorrência, o banco ficou constituído por 4.195 palavras. Após leitura pelos escolares, foram excluídas palavras que contemplavam os critérios de exclusão e que geravam desconforto por parte dos alunos. O banco ficou constituído por 4190 palavras, divididas em frequência: baixa (n= 3735, 88,59%), média (n= 374, 8,93%) e alta (n= 81, 1,93%), denominado E-LEITURA I. **Conclusão:** a elaboração de um banco de palavras de baixa, média e alta frequência de ocorrência para servir de estímulo linguístico foi adequadamente alcançado e disponibilizado para a prática clínica e pedagógica.

Correspondence address:
Adriana Marques de Oliveira
Departamento de Fonoaudiologia, Universidade Estadual Paulista "Júlio de Mesquita Filho"
Av. Hygino Muzzi Filho, 737, Marília (SP), Brasil, CEP: 17.525-900.
E-mail: adriana.oliveira@unesp.br

Received: June 05, 2019

Accepted: June 08, 2020

Study conducted at Departamento de Fonoaudiologia, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP - Marília (SP), Brasil.

[1] Faculdade de Filosofia e Ciências, Universidade Estadual Paulista "Júlio de Mesquita Filho" - Marília (SP), Brasil.

[2] Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo – USP - Ribeirão Preto (SP), Brasil.

Financial support: Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, processo nº150687/2017-6 – Pós-Doutorado Júnior – PDJ.

Conflict of interests: nothing to declare.

INTRODUCTION

The National Literacy Assessment (NLA) carried out by the Ministry of Education – MEC assesses, according to proficiency levels, the basic reading, writing, and mathematics skills by the end of the third year of elementary school. The latest results for 3rd-grade students (at the time, considered the last year of the literacy cycle) were published in 2016: The number of students considered proficient in reading was 45.3%, which means that 65.3% of 3rd-grade students had not reached the minimum reading skills to complete the literacy cycle[1].

The BEES (Basic Education Evaluation System)[†], another periodic assessment system conducted by National Institute for Space Research (Inep) on a large scale, targeted students in the 5th and 9th grades of elementary school and 3rd year of High School in 2017. In the Brazilian Portuguese language test, 39.3% of students in the 5th grade of elementary school (the last year of this cycle) were below the national average (214.5 — level 4 of the proficiency scale). This means that these students were not able, for example, to locate the information explicit in the texts, identify the main subject or the characters, and infer the meaning of words[2].

From the data presented, it becomes clear that the Brazilian education gap begins in elementary school, especially at the most important stage — literacy. The inability to read accurately and fluently directly affects reading comprehension and brings negative consequences that go beyond success in formal education, since it has the power to harm the professional and even social life of the individual, who will not be able to understand even a basic social network text[3-5].

Researchers[6-12] persistently seek to develop assessment and intervention tools for this school age group to promote and facilitate the literacy process and to identify early on if anything in the process is not evolving as expected, allowing early intervention[4]. However, in Brazil, professionals often face difficulties in developing these procedures due to the lack of words databases for selecting appropriate linguistic stimuli to evaluate and intervene, especially when it is necessary to classify them by their frequency of occurrence.

This classification, however, may involve difficulties in recognizing that the frequency of a word in written material is usually different from the frequency of the same word as observed in oral communication. Some professionals start with what is more familiar to them when classifying words as high or low frequency. As pointed out by Pinheiro[6], exposure to the spoken word affects its auditory recognition, not visual recognition. The same is true for reading, that is, when reading a word, visual recognition occurs, not auditory. The visual and auditory recognition systems are different.

In other countries, researchers maintain word frequency databases or dictionaries in which words are classified by frequency and length, from which professionals can select those that best meet their objectives[13-15]. In Spain, for example, the Real Academia Española makes available online a list of words according to their frequency.

The Brazilian Corpus[16], a collection of spoken and written texts comprises approximately one billion words with information on frequency of occurrence and categories. In this database, all grammatical categories of words are included — from verbs, whether in the infinitive, participle or conjugated form, to adverbs, prepositions, pronouns, acronyms, proper nouns, numerals, among others. As the Corpus seeks to cover all of the linguistic varieties of the Portuguese language, it does not allow the classification of words by level of schooling.

In 2007, researchers undertook a quantitative analysis of the frequency of phonemes and syllable structures in Portuguese. To do so, they created a written language corpus using the Houaiss dictionary of Portuguese. This corpus did not include abbreviations, acronyms, foreign words, homonyms, hyphenated words and those formed by juxtaposition, neither did it consider the frequency of use of words[17].

In Brazil there are lists of words, classified by frequency, orthographic regularity and length. The best known and most widely used is Pinheiro's List of Words and Pseudowords[6], which consists of 96 real words and 96 non-words.

In view of the above, there is a need to build and make available a words database classified according to the frequency of their occurrence in written material, so that professionals can build their own lists of words or use them in intervention tools.

The selection of words classified by frequency and according to level of schooling is also important to develop procedures to evaluate and intervene in metalinguistic, vocabulary, and writing abilities[6,7,18-20]. Even if the words are not used in their written form, authors start from their selection to elaborate stimuli, whether the representation is oral, written or pictorial.

Therefore, this study aims to present the process of developing a words database adequate to the reading proficiency level of elementary school students, categorized by frequency of occurrence in the presentation of written material.

METHOD

Applied research for the development of a words database appropriate for the reading proficiency level of elementary school students, to be called E-LEITURA I [e-READING I]. The applied research aims to generate knowledge for practical application to solve the problems identified[21].

Ethical Procedures

The study was registered with Plataforma Brasil (CAAE: 74853317.3.0000.5406) and approved by the Research Ethics Committee of Faculdade de Filosofia e Ciências da Universidade

[†] Since 2019, the National Literacy Assessment (ANA), the National Assessment of Basic Education (Aneb) and the National Assessment of School Performance (Anresc, also known as *Prova Brasil)* no longer exist under those names. All assessments are now called SAEB and are differentiated by the school year assessed in order to make a diagnosis of the Brazilian elementary and high school education and of the factors that may interfere in school performance. SAEB was restructured to adapt to the Common National Curricular Base (BNCC), which became the reference in the formulation of questions of the tests in Portuguese language (Brazilian variant), mathematics, natural sciences and humanities. Its application was in 2019 and the results have not been published so far[2].

Estadual Paulista Júlio de Mesquita Filho – UNESP (Protocol no. 2.375.716).

Elaboration of the Word database for students at Elementary School I

Selecting the didactic material

To prepare the database, words were taken from the courseware used by the municipal and state education systems for teaching Portuguese language in Elementary School, cycle I, in the states of São Paulo, Paraná, Rio de Janeiro and Minas Gerais.

Teaching materials were selected from three schools in the interior of the state of São Paulo, two of which were municipal schools and one state school. In the other states, only one municipal school from the capital of each state was selected. Seven collections of teaching materials distributed among the states were consulted (Tables 1 and 2).

Except for the books of the *Ler e Escrever* [Read and Write] collection, prepared by the Secretariat of Education of the State of São Paulo, the other teaching materials are organized by publishers and approved by MEC — the Ministry of Education and Culture — and integrate the National Textbook Plan — PNLD for 2016, 2017 and 2018.

Schools are free to choose those that best meet their objectives from the list of textbooks authorized by MEC. In the municipal schools consulted in the interior of the state of São Paulo, coordinators reported that, in addition to the textbooks authorized and chosen by MEC, there is a set of materials used to organize the classes that is provided by the São Paulo State Education Department, that is, the books from the *Ler e Escrever* collection.

The *Ler e Escrever* collection is adopted in the state school consulted about the teaching material. This work has specific activity and remedial make-up workbooks for each school year, as well as a single text-only workbook. For this database, the words from the remedial make-up workbooks were not typed, considering that not all students perform these activities.

Drawing up of E-LEITURA I

The methodological aspects adopted for the preparation of this study are similar to those described by Oliveira and Capellini[19,20].

Table 1. Distribution of Portuguese language teaching materials for Elementary School, Cycle I, state of São Paulo

School 1	References
Ler e escrever: coletânea de atividades – 1º ao 5º ano [*Reading and writing: collection of activities – 1st to 5th grade*]	São Paulo (State) Secretariat of Education. *Ler e escrever: coletânea de atividades – 1º ano*. Secretariat of Education. 4th. Ed. Revised and updated. São Paulo: FDE; 2014.
	São Paulo (State) Secretariat of Education. *Ler e escrever: coletânea de atividades – 2º ano*. Secretariat of Education. 7th. Ed. Revised and updated .São Paulo: FDE; 2014.
	São Paulo (State) Secretariat of Education. *Ler e escrever: coletânea de atividades – 3º ano*. Secretariat of Education. 7th. Ed. Revised and updated..São Paulo: FDE; 2014.
	São Paulo (State) Secretariat of Education. *Ler e escrever: coletânea de atividades – 4º ano*. Secretariat of Education. 6th Ed. Revised and updated..São Paulo: FDE; 2015.
	São Paulo (State) Secretariat of Education. *Ler e escrever: coletânea de atividades – 5º ano*. Secretariat of Education. 6th Ed. Revised and updated..São Paulo: FDE; 2015.
Ler e escrever: livro de textos do aluno – 1º ao 5º ano [*Reading and writing: student textbook – 1st to 5th grade*]	São Paulo (State) Secretariat of Education. *Ler e escrever: livro de textos do aluno*. Secretariat of Education. 7th Ed. São Paulo: FDE; 2013.
School 2	
Coleção Porta Aberta – Edição renovada – Letramento e alfabetização, 1º ao 5º ano [*Collection Porta Aberta –Renewed edition – Reading and Literacy, 1st to 5th grades*]	Carpaneda IPM. *Porta Aberta – língua portuguesa, 1º ano: ensino fundamental: anos iniciais*. 1st. Ed. São Paulo: FTD; 2014.
	Carpaneda IPM. *Porta Aberta – língua portuguesa, 2º ano: ensino fundamental: anos iniciais*. 1st. Ed. São Paulo: FTD; 2014.
	Carpaneda IPM. *Porta Aberta – língua portuguesa, 3º ano: ensino fundamental: anos iniciais*. 1st. Ed. São Paulo: FTD; 2014.
	Carpaneda IPM. *Porta Aberta – língua portuguesa, 4º ano: ensino fundamental: anos iniciais*. 1st. Ed. São Paulo: FTD; 2014.
	Carpaneda IPM. *Porta Aberta – língua portuguesa, 5º ano: ensino fundamental: anos iniciais*. 1st. Ed. São Paulo: FTD; 2014.
School 3	
Aprender e criar – Letramento e Alfabetização 1º ao 3º ano [*Aprender e criar – Reading and Literacy – 1st to 3rd years*]	Neves AAA, Carvalho A, Bevilacqua E, Grilo M. *Aprender e criar: letramento e alfabetização, 1*, 1st. 2nd Ed. São Paulo: Escala Educacional; 2014.
	Neves AAA, Carvalho A, Bevilacqua E, Grilo M. *Aprender e criar: letramento e alfabetização, 2*. 2nd Ed. São Paulo: Escala Educacional; 2014.
	Neves AAA, Carvalho A, Bevilacqua E, Grilo M. Aprender e criar: letramento e alfabetização, 3. 2nd Ed. São Paulo: Escala Educacional; 2014.
Ápis: Língua Portuguesa – 4º ao 5º ano [Ápis*: Portuguese Language – 4th to 5th grade*]	Borgatto AMT, Bertin TCH, Marchezi VLC. Ápis*: Língua Portuguesa 4º ano*. 2nd Ed. São Paulo: Ática; 2014.
	Borgatto AMT, Bertin TCH, Marchezi VLC. Ápis*: Língua Portuguesa 5º ano*. 2nd Ed. São Paulo: Ática; 2014.

Table 2. Distribution of Portuguese language teaching materials for Elementary School, Cycle I, states of Minas Gerais, Paraná and Rio de Janeiro

Minas Gerais	References
Coleção Quatro Cantos – Português – Letramento e Alfabetização, 1º ao 3º ano	Porto A, Antoniol V. Coleção Quatro Cantos: Português – Letramento e Alfabetização – 1º ano: livro do aluno. Belo Horizonte: Dimensão; 2013.
	Porto A, Antoniol V. Coleção Quatro Cantos: Português – Letramento e Alfabetização – 2º ano: livro do aluno. Belo Horizonte: Dimensão; 2013.
	Porto A, Antoniol V. Coleção Quatro Cantos: Português – Letramento e Alfabetização – 3º ano: livro do aluno. Belo Horizonte: Dimensão; 2013.
Projeto Buriti: português: Ensino Fundamental: anos iniciais, 4º ao 5º ano	Editora Moderna. Projeto Buriti: português: ensino fundamental anos iniciais – 4º ano. 3ª Ed. São Paulo: Moderna; 2014.
	Editora Moderna. Projeto Buriti: português: ensino fundamental anos iniciais – 5º ano. 3ª Ed. São Paulo: Moderna; 2014.
Paraná	
Aprender juntos – Letramento e Alfabetização – 1º ao 5º ano	Vasconcelos A. Aprender Juntos: letramento e alfabetização, 1º ano: ensino fundamental: anos iniciais. 4ª Ed. São Paulo: Edições SM; 2014.
	Vasconcelos A. Aprender Juntos: letramento e alfabetização, 2º ano: ensino fundamental: anos iniciais. 4ª Ed. São Paulo: Edições SM; 2014.
	Vasconcelos A. Aprender Juntos: letramento e alfabetização, 3º ano: ensino fundamental: anos iniciais. 4ª Ed. São Paulo: Edições SM; 2014.
	Vasconcelos A. Aprender Juntos: português, 4º ano: ensino fundamental: anos iniciais. 4ª Ed. São Paulo: Edições SM; 2014.
	Vasconcelos A. Aprender Juntos: português, 5º ano: ensino fundamental: anos iniciais. 4ª Ed. São Paulo: Edições SM; 2014.
Rio de Janeiro	
Aprender juntos – Letramento e Alfabetização – 1º ao 3º ano	Vasconcelos A. Aprender Juntos: letramento e alfabetização, 1º ano: ensino fundamental: anos iniciais. 4ª Ed. São Paulo: Edições SM; 2014.
	Vasconcelos A. Aprender Juntos: letramento e alfabetização, 2º ano: ensino fundamental: anos iniciais. 4ª Ed. São Paulo: Edições SM; 2014.
	Vasconcelos A. Aprender Juntos: letramento e alfabetização, 3º ano: ensino fundamental: anos iniciais. 4ª Ed. São Paulo: Edições SM; 2014.
Projeto Buriti: português: Ensino Fundamental: anos iniciais, 4º ao 5º ano	Editora Moderna. Projeto Buriti: português: ensino fundamental anos iniciais – 4º ano. 3ª Ed. São Paulo: Moderna; 2014.
	Editora Moderna. Projeto Buriti: português: ensino fundamental anos iniciais – 5º ano. 3ª Ed. São Paulo: Moderna; 2014.

All words in the texts integrating the didactic materials were typed into an Excel spreadsheet. After typing, those that were in accordance with the inclusion criteria of E-LEITURA I were selected.

To select the words, we chose to type those from the texts in the Portuguese textbooks. Words belonging to both the noun and adjective classes were inserted, according to the context in which they are used, that is, the words that could be classified as nouns and/or adjectives (classification contained in the Michaelis online Portuguese dictionary).

It should be noted that it was not enough for the adjective to be a noun; both classifications, adjective and noun, needed to be present in the Portuguese dictionary. For example, in the case of words such as *coruja* [owl], included in the dictionaries as both noun and adjective, as in "She is a *mãe coruja*" [i.e., a "mama bear"] (adjective) and "*Corujas* [owls] do not usually hunt during the day" (noun). The word "*abacate*" [avocado] is classified only as a noun in the dictionaries, which is why both these words were included in the words database.

Nouns were selected because this class is frequent in any text, as they have important syntactic functions in sentences. We chose to stick to this class of words because it is the core of the noun phrase[19,20].

Nouns can be recognized by some criteria, such as those cited by Cegalla[22]:

1. They stand for the names of beings;
2. They are always nuclei of the nominal syntagma;
3. They generally accept an article;
4. Syntactically, they can have several functions: subject, verbal complement, passive agent, nominal complement, predicative, appositive;
5. They inflect for gender, number and degree.

The nouns and adjectives that are inflected for gender, number and degree (augmentative/diminutive) and that have this classification in the dictionaries were included in the words database.

Homophones that could present ambiguity depending on the context were excluded from the database because they are isolated words, i.e., their require the sentence to retrieve their meaning and pronunciation. These are homonymous homographs (written the same way, decoded differently) and perfect homonyms (written the same way and pronounced equally).

Also excluded from the words database were words written in other languages (even those already incorporated into the Portuguese dictionary), abbreviations, adverbs, adverbial and prepositional phrases, adjectives, names of months of the year, numerals, and words in the augmentative or diminutive, in

addition to slang terms and words composed by juxtaposition. Also not considered were proper nouns, words with spelling recorded incorrectly in the material and/or used to represent popular pronunciation, words with divergent spelling and archaisms, i.e., words no longer in use.

Since in Brazilian Portuguese the dominant gender is masculine, feminine words were changed to the masculine. Words were kept in the feminine if there was another term for the masculine or if the suffix was modified.

As to number, the singular was always used. Words that, when changed from plural to singular assumed a homonymous homograph form or became a perfect homonym, or even if it presented any kind of ambiguity, were excluded from the database.

Following the spelling reform, the umlaut is no longer used (e.g., the words *bilíngue, pinguim, antiguidade* previously spelled as *bilíngüe, pingüim, antigüidade*) and open diphthongs (*ei, oi, eu*) are only accented at the end of the word. Words written according with the old spelling rules, used before the spelling reform, were adapted.

When preparing E-LEITURA II and III[19,20], some words that could cause discomfort or invite jokes from students were removed, in addition to words that could cause confusion when reading, such as the *face* [fˈa.si], that was pronounced as in English, *face* [fˈej.si] — from *Facebook*. Thus, some words were also excluded from the list, since they could arouse unwanted behaviors in elementary school students, such as *face, sexo, sexual, calcinha, cueca, virgem* e *capeta* [face, sex, sexual, panties, underpants, virgin, and devil].

After application of these selection criteria, all words and the number of times they appeared in the material were counted to determine the frequency of occurrence in each school year. The words were organized in a single database and sent to the statistician to determine the low, medium, and high frequency words common to all the years.

From the above-mentioned procedures, a unique words database was created for Elementary School Level I (1st to 5th grades), named E-LEITURA I. The selected words were classified according to their frequency of occurrence in the didactic material consulted and, for this, the tertiles of the distribution, the average frequency and the cutoff point of the tertiles — according to the frequencies close to the center — were used to rank them as low, medium and high frequency.

Participants

To detect possible flaws in the selection of words (those that did not meet the inclusion and exclusion criteria, were typed incorrectly, or that elicited reading refusal behavior), 50 students from a state public school of a city in center-western São Paulo participated in this study.

Parents or legal guardians signed two copies of the Informed Consent Form (ICF), according to National Health Council resolution CNS 196/96. The 50 participants were grouped as follows:

- GI: 10 schoolchildren from the 1st year of Elementary School cycle I;
- GII: 10 students from the 2nd year of Elementary School Cycle I;
- GIII: 10 students from the 3rd year of Basic Education Cycle I;
- GIV: 10 students from the 4th year of Basic Education Cycle I;
- GV: 10 students from the 5th year of the Elementary School Cycle I.

The participating students were selected according to inclusion and exclusion criteria. Information was obtained from school records and/or the teachers.

Inclusion criteria for selecting participants were: 1) being regularly enrolled in the Elementary School Cycle I; 2) parents or guardians signing the Free and Informed Consent Form; 3) signing the Consent Form. The study excluded students: 1) who refused to participate, although their parents or guardians had signed the consent form; 2) with interdisciplinary diagnosis of learning disability, dyslexia, or attention deficit hyperactivity disorder; 3) with learning complaints; 4) with language or speech impairments; 5) with visual and auditory impairment; 6) with diagnosis of genetic or neurological syndromes; 7) with a history of repetition; and 8) intellectually impaired students.

Procedures

The low, medium, and high frequency words of E-LEITURA I were presented to the students on a sheet of A4 paper, typed in Arial size 14 font, capital letters, double spacing, separated into three columns. Each sheet had, on average, 51 words, which were read aloud individually, one at a time, by the schoolchildren.

Each student was asked to read as many words as he or she could for 20 minutes. When the time was up, they would wait until they finished the page, they were on to end the assessment. The next student would start reading from the point where the last one had stopped.

The procedure initially called for students to read all the words in the database in 15-minute sessions. It was estimated that an average of 15 sessions would be necessary. However, this procedure had to be redesigned on the first day of data collection, since the students showed fatigue. In view of this, data collection was reduced to a maximum of two sessions, each lasting 20 minutes.

Second and third graders participated twice in the reading of the words database, up to 20 minutes each time, with a two-week interval between sessions. This was done randomly. When the 10 students had finished reading, they started over with the first student until they were finished reading the database. As a result, all of the students participated again. Fourth and fifth graders, on the other hand, participated only once. The students were taken from the classroom at the school's convenience, after authorization from the teacher and the principal.

Fourth and fifth graders read the words in August and September 2nd and 3rd graders in October and November, and 1st graders in November. Reading by 1st grade students was the last to be started, in order to give them more time for literacy classes and, therefore, better conditions to receive instructions regarding the decoding activity.

Table 3. Description of the number and frequency of words typed for the elaboration of E-LEITURA I

COLLECTION	Number of words typed	Frequency of words
Ler e escrever: coletânea de atividades 1º ao 5º ano	20,446	104,199
Coleção Porta Aberta – Letramento e alfabetização 1º ao 5º ano	8,992	43,335
Aprender juntos – Letramento e Alfabetização – 1º ao 5º ano	10,555	45,596
Aprender e criar – Letramento e Alfabetização 1º ao 3º ano	3,573	13,219
Coleção Quatro Cantos – Português – Letramento e Alfabetização, 1º ao 3º ano	3,651	11,724
Ápis: Língua Portuguesa – 4º ao 5º ano	9,830	49,834
Projeto Buriti: português: Ensino Fundamental: anos iniciais – 4º ao 5º ano	4,869	18,383
TOTAL	61,916	286,290

Table 4. Distribution from the cut-off point of tertiles for determining the frequency of occurrence of words, total frequencies and average frequency, number of words per frequency

	E-LEITURA I
Total frequencies	48,185
Average frequency	11.48
Cutoff point	
1st tertile	22/23
2nd tertile	100/101
Number of repetitions per frequency	
Low	1-22
Medium	23-100
High	101-590
Number of words per frequency	
Low	3,738
Medium	375
High	82
Total of words in E-LEITURA I	4,195

Analysis of the results

The information collected was recorded in a Microsoft Excel program database. The tertiles of the distribution were used to classify the words as low, medium, and high frequency, as well as the average frequency and cut-off point of the tertiles (based on the frequencies that were close to the center). The tertile divides the interval of a frequency distribution into three classes of equal number (33.33%). In a symmetrical distribution, the values of the tertiles are found checking which ones are in the interval between 33.33% and 66.66%.

RESULTS

Description of the results of the word selection

286,290 words were typed (counted by frequency of occurrence, i.e., the same word appears more than once) among articles, prepositions, adjectives, verbs, nouns and others. Removing the repeated words, 61,916 words remained per collection of books used for the preparation of E-LEITURA I, as presented in Table 3.

All 61,916 words were analyzed and selected according to the established criteria. After analysis and selection, E-LEITURA I consisted of 4,195 words submitted for classification by frequency of occurrence (low, medium, or high).

The values of the cumulative distribution (cumulative frequency) were used in E-LEITURA I because it is a words database with many repetitions. Data distribution, therefore, is asymmetric. Since these are nominal variables, the distribution of tertiles in E-LEITURA I does not contain exactly 33.33% of the total.

In this case, the first tertile starts at repetition 22, cumulative percentage 33.30%, until the 34.07% percentage, and ends at repetition 23, cumulative percentage 34.12%. The second tertile starts at cumulative percentage 66.56% (repetition 100) and ends at cumulative percentage 66.77% (repetition 101).

From the tertile values, it was possible to classify the words as low, medium, and high frequency, as shown in Table 4.

Canonical syllables (consonant "C" and vowel "V" – CV) represent 44.25% of the total words in E-LEITURA I, or 1,844 of them, followed by non-canonical CVC (n=855, 20.52%, example: "*bactéria*"), VC (n=458, 10.99%, as in "*ar*"), V (n=400, 9.60%, as in "*idade*") and CCV (n=321, 7.70%, as in "*brejo*"). Below five percent are the following: CVV (n = 136, 3.26%, as in "*goiaba*"), CCVC (n= 63, 1.51%, as in "*clássico*"), VV (n= 25, 0.60%, as in "*autor*"), CCVV (n= 15, 0.36%, as in "*chão*"), CVCC (n= 13, 0.31%, as in "*monstro*"), CVVC (n= 12, 0.29%, as in "*questão*"), VVC (n= 1, 0.26%, "*Austríaco*"), CCVCC (n= 10, 0.24%, as in "*translação*"), CVVC (n= 2, 0.05%, as in "*cais*"), VVC (N=1, 0.02%) and CCVVC (n= 1, 0.02%, as in "*braille*"). In all, 16 syllabic structures were found.

As for length, trisyllabic words are the ones with the highest occurrence, representing 35.23% or 1,468 words, followed by polysyllabic ones with four syllables, with 1,159 words (27.81%). Disyllabic words and polysyllabic words with five syllables come next with 87 (19.85%) and 509 (12.22%) words, respectively. Polysyllabic words with more than six syllables represent 3.62% of the total E-LEITURA I words, being n= 115 (2.76%) with six syllables, n= 31 (0.74%) with seven, n= 4 (0.10%) with eight syllables and n= 1 (0.02%) with ten syllables. Monosyllables make up 1.27% of the database, with 53 words.

The distribution of low, medium and high frequency words by length and syllabic complexity is shown in Table 5.

Description of the E-LEITURA I reading results

It was not possible to evaluate the 1st-grade students. When we started reading with these children, while still dealing with high frequency words, we noticed that they only read monosyllabic and canonical disyllabic (CV) words. For example, if the words had the letter x, as in "*bruxa"*, or the letter r in the middle of the

Table 5. Distribution of high, medium, and low frequency words according to syllable length and initial syllable complexity

E-LEITURA I					
High frequency					
Syllable length			Complexity of Initial Syllable		
Monosyllable	13 (16,05%)	V	7 (8,64%)	CVC	12 (14,81%)
Disyllable	42 (51,85%)	VC	5 (6,17%)	CVV	8 (9,88%)
Trisyllable	23 (28,40%)	CV	40 (49,38%)	CCVC	4 (4,94%)
Polysyllable (four syllables)	3 (3,70%)	CCV	4 (4,94%)	CCVV	1 (1,23%)
Total words	81 (100%)				
Average Frequency					
Syllable length			Complexity of Initial Syllable		
Monosyllable	18 (4,81%)	V	31 (8,29%)	VCC	1 (0,27%)
Disyllable	145 (38,77%)	VC	32 (8,56%)	CCVC	8 (2,14%)
Trisyllable	140 (37,43%)	CV	180 (48,13%)	CCVV	2 (0,53%)
Polysyllable (four syllables)	62 (16,58%)	CCV	21 (5,61%)	CVVC	3 (0,80%)
Polysyllable (five syllables)	7 (1,87%)	CVC	79 (21,12%)	CVCC	1 (0,27%)
polysyllable (six syllables)	2 (0,53%)	CVV	16 (4,28%)		
Total words	374 (100%)				
Low frequency					
Syllable length			Complexity of Initial Syllable		
Monosyllable	22 (0,59%)	V	362 (9,75%)	CVVV	2 (0,05%)
Disyllable	642 (17,24%)	VC	421 (11,34%)	CVCC	12 (0,32%)
Trisyllable	1307 (35,16%)	CV	1626 (43,75%)	CCVC	51 (1,37%)
Polysyllable (four syllables)	1094 (29,47%)	VV	26 (0,67%)	CCVV	12 (0,32%)
Polysyllable (five syllables)	502 (13,52%)	VVC	1 (0,03%)	CCVCC	10 (0,27)
polysyllable (six syllables)	113 (3,04%)	VCC	10 (0,27%)	CCVVC	1 (0,03%)
polysyllable (seven syllables)	31 (0,84%)	CCV	296 (7,97%)		
polysyllable (eight syllables)	4 (0,11%)	CVC	764 (20,58%)		
polysyllable (nine syllables)	0 (0,00%)	CVV	112 (3,02%)		
polysyllable (ten syllables)	1 (0,03%)	CVVC	9 (0,24%)		
Total words	3716 (100%)				
Total of words in E-LEITURA I	4190				

word, as in "*porta*", or at the end, as in "*amor*", they expressed that they could not read them. Two 1st-grade students were called to read the high-frequency words; afterwards, the decision to exclude the 1st grade from the study was made given that as the database was too complex for these target children.

After the words were read by the students from the 2nd to the 5th grades, we noticed that some of the words met the exclusion criteria and others generated discomfort on the part of the students. Therefore, the following words were removed from the database: high-frequency: *cor* [color/"by heart"] (which can be read with the tonic vowel open or closed / ɔ/ or /o/ depending on the meaning); medium-frequency: *ministro* (from the verb *ministrar* [1st person singular, present tense, of "to minister"]) and the low-frequency words: *arrear* (verb), *palhava* (not found in the Michaelis dictionary and in the VOLP — Orthographic Vocabulary of Portuguese of the Brazilian Academy of Letters) and the word *inferno* [hell] (refusal to read).

After exclusion of these words, E-LEITURA I consisted of 4,190 words, divided according to frequency: high — 81 words, which corresponds to 1.93% of the database; medium — 374 (8.93%) words; and low — 3,735 (88.59)% words.

The E-LEITURA I words database is presented in Supplementary Table 1, 2 and 3, broken down into parts — Supplementary Table 1 has high-frequency words; Supplementary Table 2 has medium-frequency words, and Supplementary Table 3 has low-frequency words. Words appear in alphabetical order, indicating the number of syllables (syllable length), word complexity according to the initial syllable structure (how consonants and vowels are organized in the syllable; vowels are represented by the letter "V" and consonants by the letter "C"), as well as the number of repetitions of each word in the database.

DISCUSSION

Development of the E-LEITURA I words database was based on the need for linguistic stimuli to create assessment and intervention tools for elementary school students. When using words for reading assessment, for example, the isolated word list is one of the most effective materials. However, the selection of words to compose a list should follow some criteria according to psycholinguistic characteristics such as regularity, length, and frequency of occurrence[6,18,23]. When we talk about frequency of occurrence, however, we are not referring to the oral experience of the student with that word, but to the written mode, the reading experience, the number of times they have visualized this word, that is, the visual representation of the word[6]. This attention in the selection by frequency and school level is important not only for reading activities, but also for

the development of instruments to assess and intervene in metalinguistic skills, vocabulary, and writing[6,7,19,20].

For E-LEITURA I, we chose to use the material prepared by the Secretariat of Education of the State of São Paulo and the one recommended by MEC (at a national level) to be worked on in 2017, both in São Paulo and in other states (Paraná, Minas Gerais and Rio de Janeiro) for public and municipal schools. There are no specific books for each region of the country, therefore regionalisms are not privileged.

Not considering regionalisms is the main criticism made to the few existing words databases[19,20]. In this study, it was observed that, despite the larger number of books used, no significant increase of words in relation to known databases were observed[19,20], but rather an increase in the number of times each word appears.

In E-LEITURA I, the CV structure is predominant, followed by the CVC structure. This finding is in line with studies that surveyed the profile of Brazilian Portuguese words and identified a higher occurrence of the CV structure followed by CVC[17,24]. The CV (canonical) structure is the most frequent in most languages, including Portuguese, followed by the non-canonical V, VC and CVC structures. Being the predominant one, the CV structure is among the first to emerge in the acquisition of the linguistic system and schoolchildren tend to learn this syllabic structure first[17,24,25].

According to the guidelines of the Common National Curricular Base — BNCC[26] — the CV, V, CVC and CCV structures should be acquired throughout the 2nd and 3rd grades, when students should be able to correctly read and write words with such structures in all syllables. The VC structure should be mastered by the 3rd grade, and the VV and CVV structures throughout the 3rd and 4th grades.

When classifying which structures occur more frequently in the database (CV, CVC, VC, V, CCV, CVV, CCVC, and VV), it is noticeable that they follow the development pattern of the BNCC guidelines[26], which shows that E-LEITURA is in line with what is specified for Elementary I. The more complex CCVC structure is present in the sequence, probably because E-LEITURA consists of the teaching material from 1st to 5th grades of Elementary I.

Regarding word length, the findings of this study are also in line with the literature, which states that Portuguese is a predominantly trisyllabic language, with few monosyllables[17,25]. However, contrary to what occurs in Marques' analysis[25], in which a higher number of trisyllabic words appear in first place followed by disyllabic words, this study found a higher number of trisyllabic words followed by polysyllabic words with four syllables before disyllabic words.

In this study the lowest number of syllables found was one and the highest was 10, and the word with the highest number was "*otorrinolaringologista*" [otorhinolaryngologist]. A study carried out in Brazil[17] that complied a written language corpus from the Houaiss dictionary found that the smallest number of syllables was one and the largest word had 20 syllables ("*pneumoultramicroscopicossilicovulcanoconiótico*" [pneumoultramicroscopicossilicovulcanoconiotic]).

The 1st-grade students were not able to read the words from the database, as they only read the canonical monosyllabic and disyllabic words (CV). Words with CVC or CVV structures already on the first syllable could not be read by them. Such behavior is in accordance with the Common National Curriculum Base – BNCC[26], which states that decoding and reading fluency — defined in the document as the reading of new words and words of frequent use — will be achieved globally by memorization. According to the Common National Curricular Base document, only students who already understand the writing system have this ability, which can take place by the end of the 2nd grade.

The schoolchildren's behavior to stop reading, reporting tiredness, and the decision to limit the time reflect the fact that reading, depending on the route used, demands more phonological or visuospatial working memory, which possibly causes overload[27-30]. There is a consensus in the literature that these words require more reading time than high- and medium-frequency words, and that the number of errors also usually increases[6,7,18-20,27-29]. Since they are low-frequency words, the reader does not have their representation in their input visual lexicon, which leads them to read phonologically, thus requiring their attention more frequently.

The main objective of the words database is to offer a tool from which professionals can obtain the linguistic stimuli needed to select words according to their frequency of occurrence in the written material.

CONCLUDING REMARKS

The objective of developing a database of low-, medium- and high-frequency words from the written material to serve as a linguistic stimulus was adequately achieved.

Of the total number of words in E-LEITURA I, low-frequency words correspond to 88.59%. As for the initial syllable structure, canonical syllables are the most frequent. Regarding length, trisyllables are the ones with the highest occurrence, followed by polysyllables with four syllables.

Although named E-LEITURA [e-READING], the words database was created based on the frequency of occurrence of the words, so these stimuli can also be used in writing and spelling activities. The frequency of the words does not change if the object of analysis is decoding or coding.

Access to the words database and the possibilities of tools that can be developed from the linguistic stimuli presented therein may help professionals to identify and intervene early in cases of reading difficulties.

ACKNOWLEDGEMENTS

To Gabriela Franco dos Santos Liporaci, Irene B. Marques de Oliveira and Larissa Sellin for their assistance in typing the words for the elaboration of E-LEITURA I. To Cristiane Tomazinho Fumagalli, Edmilton Oseias da Cunha, Fernanda Boatto Belgamasco, Juliana Mendes Alves, Luciana Cássia Pereira Capel, Luciana Cordeiro Felipetto and Patrícia Hiraoka for providing and recommending the didactic material.

REFERENCES

1. INEP: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. MEC: Ministério da Educação. Sistema de avaliação da Educação Básica – avaliação nacional da alfabetização – Edição 2016 [Internet]. Brasília: INEP/MEC; 2017 [citado em 2017 Out 9]. Disponível em: http://portal.mec.gov.br/docman/outubro-2017-pdf/75181-resultados-ana-2016-pdf/file
2. INEP: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Relatório SAEB. Brasília: INEP/MEC; 2019.
3. Lervåg A, Hulme C, Melby-Lervåg M. Unpicking the developmental relationship between oral language skills and reading comprehension: it's simple, but complex. Child Dev. 2018;89(5):1821-38. http://dx.doi.org/10.1111/cdev.12861. PMid:28605008.
4. Colenbrander D, Ricketts J, Breadmore HL. Early identification of Dyslexia: understanding the issues. Lang Speech Hear Serv Sch. 2018;49(4):817-28. http://dx.doi.org/10.1044/2018_LSHSS-DYSLC-18-0007. PMid:30458543.
5. Oakhill J, Cain K, Elbro C. Compreensão de leitura: teoria e prática. São Paulo: Hogrefe CETEPP; 2017.
6. Pinheiro AMV. Leitura e escrita: uma abordagem cognitiva. 2. ed. Campinas: Livro Pleno; 2006.
7. Capellini SA, Oliveira AM, Cuetos F. PROLEC: provas de avaliação dos processos de leitura. 3. ed. São Paulo: Casa do Psicólogo; 2014.
8. Salles JF, Parente MAPP. Avaliação da leitura e escrita de palavras em crianças de 2ª série: abordagem neuropsicológica cognitiva. Psicol Reflex Crit. 2007;20(2):220-8. http://dx.doi.org/10.1590/S0102-79722007000200007.
9. Lamprecht R, Santos RM, Freitas GM, Brodacz R, Siqueira M, Costa AC, et al. Confias: consciência fonológica instrumento de avaliação sequência. São Paulo: Pearson; 2015.
10. Santos B, Capellini SA. PRONAR-LE - Programa de remediação com nomeação automática rápida e leitura. Ribeirão Preto: Booktoy; 2018.
11. Seabra AG, Capovilla FC. TCLPP - Teste de competência de leitura de palavras e pseudopalavras. São Paulo (SP): Memnon; 2010.
12. Silva C, Capellini SA. Programa de Intervenção fonológica para escolares em fase inicial de alfabetização: manual e caderno do aplicador. Ribeirão Preto: Booktoy; 2019.
13. Balota DA, Yap MJ, Cortese MJ, Hutchison KA, Kessler B, Loftis B, et al. The English Lexicon Project. Behav Res Methods. 2007;39(3):445-59. http://dx.doi.org/10.3758/BF03193014. PMid:17958156.
14. Martínez JA, García E. Diccionario. Frecuencias del castellano escrito en niños de 6 a 12 años. Salamanca: Publicaciones de la Universidad Pontificia, 2004.
15. López MXB. O Galego fundamental: dicionário de frecuencias. Santiago de Compostela: A Coruña Fundación Pedro Barrié de la Maza; 2007.
16. Corpus Brasileiro. Projeto AC/DC: corpo Corpus Brasileiro [Internet]. 2018 [citado em 2018 Nov 11]. Disponível em: http://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS
17. Viaro ME, Guimarães-Filho ZO. Análise quantitativa da frequência dos fonemas e estruturas silábicas portuguesas. Estudos Linguísticos. 2007;26(1):27-36.
18. Cuetos F. Psicología de la lectura. 8. ed. Las Rozas, Madrid: Wolters Kluwer España; 2010.
19. Oliveira AM, Capellini SA. E-LEITURA II: banco de palavras para leitura de escolares do Ensino Fundamental II. CoDAS. 2016;28(6):778-817. http://dx.doi.org/10.1590/2317-1782/20162016049. PMid:27982255.
20. Oliveira AM, Capellini SA. Banco de palavras para leitura de escolares do ensino médio: E-LEITURA III. Rev CEFAC. 2016;18(6):1404-46. http://dx.doi.org/10.1590/1982-0216201618610516.
21. Silva EL, Menezes EM. Metodologia da pesquisa e elaboração de dissertação. 4. ed. Florianópolis: UFSC; 2005.
22. Cegalla DP. Novíssima Gramática da Língua Portuguesa. 48. ed. São Paulo: Companhia Editora Nacional; 2008.
23. Pinheiro AMV, Rothe-Neves R. Avaliação cognitiva de leitura e escrita: as tarefas de leitura em voz alta e ditado. Psicol Refl Crit (Lond). 2001;14(2):399-408. http://dx.doi.org/10.1590/S0102-79722001000200014.
24. Matzenauer CLB. Bases para o entendimento da aquisição fonológica. In: Lamprecht RR. (Ed.), Aquisição Fonológica do Português: perfil de desenvolvimento e subsídios para a terapia. Porto Alegre: Artmed; 2004. p. 33-58.
25. Marques LF. Estruturas silábicas do português do Brasil: uma análise tipológica [dissertação]. São Paulo: Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo; 2008. http://dx.doi.org/10.11606/D.8.2008.tde-06082009-163245.
26. Brasil. Ministério da Educação. Secretaria da Educação Básica. Base nacional comum curricular. Brasília, DF: Ministério da Educação; 2016. [citado em 2018 Dez 11]. Disponível em: http://basenacionalcomum.mec.gov.br/#/site/inicio
27. Baddeley A. Memória de trabalho. In: Baddley A, Anderson MC, Eysenck MW. (Eds.), Memória. Porto Alegre: Artmed; 2011. p. 54-82.
28. Gonçalves HA, Viapiana VF, Sartori MS, Giacomoni CH, Stein LM, Fonseca RP. Funções executivas predizem o processamento de habilidades básicas de leitura, escrita e matemática? Neuropsicol Lat AM. 2017;9(3):42-54. http://dx.doi.org/10.5579/rnl.2016.0393.
29. Zhao J, Liu M, Liu H, Huang C. Increased deficit of visual attention span with development in Chinese children with developmental dyslexia. Sci Rep. 2018;8(1):3153. http://dx.doi.org/10.1038/s41598-018-21578-5. PMid:29453430.
30. Awadh FHR, Phénix T, Antzaka A, Lallier M, Carreiras M, Valdois S. Cross-Language Modulation of visual attention span: an Arabic-French-Spanish in skilled adult readers. Front Psychol. 2018;7:307. http://dx.doi.org/10.3389/fpsyg.2016.00307. PMid:27014125.

Authors contributions

AMO: participated in the drawing up and planning of the research project, data collection, data analysis and interpretation, writing and critical revision of the manuscript; JLFS: participated in planning of the research project, data analysis (responsible for using the frequencies observed — divided into tertiles — for classification of words) and interpretation, and critical revision of the manuscript; SAC: participated in the drawing up, planning and guidance of the research project, writing and critical revision of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material accompanies this paper.

Supplementary Table 1. Reading words database for elementary school students – E-LEITURA I – high-frequency words

Supplementary Table 2. Elementary I school students' reading words database – E-LEITURA I – medium-frequency words

Supplementary Table 3. Reading words database for elementary school students – E-LEITURAI – low-frequency words

This material is available as part of the online article from https://www.scielo.br/j/codas/