

Frequency of occurrence of words in textbooks exposed to Brazilian children in the early years of Elementary School

Ângela Maria Vieira Pinheiro¹

Abstract: The *corpus* presented here has become known, in Brazilian literature, as “Pinheiro’s word count” (1996). This collection of material was developed to allow the identification of the frequency of occurrence of words in written material exposed to children in the early years of Elementary education, in both public and private schools, in the state of Minas Gerais, Brazil, during the period 1990-1994. The examination of a large body of words (in the region of two million), taken from a wide variety of sources, permitted the identification of a vocabulary (lexicon), composed of words classified as being of high, medium or low frequency of occurrence, that is representative of the vocabulary to which children are exposed during specific periods of the process of learning to read and write. Lexicographers, linguists, psycholinguists and educators will find within this work important data that can be used for different research purposes. In the context of study concerning the learning and development of reading and writing in Brazil, the present *corpus* of words has, since its publication, been the main reference source for research in this area.

¹ Professor at the Department of Psychology, Faculdade de Filosofia e Ciências Humanas (FAFICH), Universidade Federal de Minas Gerais Universidade Federal de Minas Gerais, Brazil. Contact: pinheiroamva@gmail.com

Introduction

In 1996, Pinheiro made available to the Brazilian scientific community a *corpus* concerning the written vocabulary to which children are exposed. Given the importance of this work, which since its publication has been the reference for research in the area of reading and writing development, it is presented here on the CHILDES Platform, accompanied by updated information and materials relevant to its comprehension and use.

Brief history

The *corpus* referred to here was developed by means of two studies. The first consisted of the selection of the publications most frequently used in the curricula at pre-school and during the initial years (the 1st to 4th school years) of Elementary school in both the Public (Municipal and State schools) and private sectors in the city of Belo Horizonte (Pinheiro, 1991). This study generated the *input* for the second study, which consisted of the formation of a *corpus* of words and the subsequent counting of the frequency of occurrence of each of them (Pinheiro, 1996, 1996).

Study 1 – Selection of the inputs of the *corpus*

The work consisted of the following steps:

- 1) A survey of the number of 1st to 5th year/grade Elementary schools in the Belo Horizonte teaching networks, followed by a tiered draw of 10% of schools in each region of each network;
- 2) Development of the investigative tool – a questionnaire inquiring about the most commonly used textbooks, supplementary books and written material in the areas of Portuguese, Mathematics, Social Studies and Sciences;
- 3) Sending this questionnaire to be completed by the Educational Supervisors of the selected establishments; and
- 4) The selection of all publications that received three or more citations.

This procedure resulted in the identification of 124 books (8.19% of the 1514 different titles surveyed), the majority of these having been adopted by the Ministry of Education for use in all Brazilian States. Table 1 shows the distribution of these written materials by school year.

Table 1.

*Frequency of citation of publications in the subject areas investigated from the 1st to 5th years of Elementary school**

Subject	School year					Total
	1 st	2 nd	3 rd	4 th	5 th	
Language (Portuguese)	8	13	13	11	11	56
Mathematics	4	8	7	8	6	33
Social Studies/Science	4	6	7	9	9	35
Total	16	27	27	28	26	124

*Publications and their respective references are presented in Appendices 1 and 2.

Study 2 – Forming the *corpus* of words

The texts extracted from the publications selected in Study 1 were typed and edited (with the formal consent of their respective publishers) and then processed by the Oxford Concordance Programme (OCP) which registered 1,774,164 words (the quantity of words or *tokens* in the *corpus*) in the 124 publications, yielding a vocabulary of 40,509 words (the number of different words or *types*). Table 2 shows the distribution of this data by school year.

Table 2.

Distribution of the number of words (tokens) and vocabulary (types) by school year

School year	<i>Tokens</i>	<i>Types</i>
1 st	34045	4573
2 nd	209216	11640
3 rd	363313	18297
4 th	551506	24521
5 th	616084	28742
TOTAL	1774164	40509*

*Excludes duplicated words.

The criterion for the inclusion of an item in the *corpus* was its occurrence in terms of *graphic word type*, as the OCP could only recognise different strings of letters and characters. Therefore, this word frequency count, as per Carrol, Davis e Richman's (1971) book, gives frequencies under separate entries for *acabar*, *acaba*, *acabada*, *acabado*, *acabam*....., for instance. As Carrol (1972) explained, "the listing of frequencies for separate graphic types rather than by dictionary entries has the advantage of presenting the basic data which anyone can combine in whatever way he wishes" (p. 1072). Taking the example above, one can find the combined frequency of all words with the morpheme (radical) *acab-* as a component.

The exception from the above criterion was for those items composed of a string of graphic word types such as the onomatopoetic (e.g., *pam ram pam pam*), interjections (e.g., *ufa ufa*) and words giving a notion of melody or any invented sound (e.g., *la la la la*), in such cases the whole set was considered as a lexical unit. Items of this type were entered without spaces between their components (e.g., *lalalala*) and subsequently, in the editorial processing of the word lists generated by the OCP, the spaces were put back into their respective places.

The hyphen was retained only in compound nouns and in the cases in which its removal produced a change in the morphology of the word, as in the inter-clitical forms such as *cortá-la*, *conhecê-la*, where the accented morphemes carry both lexical and grammatical information. The diacritic indicates, thus, that the morpheme is accented and that its attachment implicates its dependency as an object pronoun. The items *cortá* and *conhecê* are not words by themselves. Following this reasoning, in grammatical forms such as *viu-se*, *comeu-se* and *deu-me* in which each of the two items joined by the hyphen are words on their own, the hyphen was removed.

Also, in the same way as Carrol, Davis and Richman (1971), words such as *banco*, *canto* e *mato* (that can have the function of a verb or common noun) were considered as undifferentiated units. Access to the grammatical and semantic complexity of these words can only be achieved by querying the context in which they occur (information not available in this *corpus*).

Finally, numerals (e.g., 0, 1, 2, 3, etc.), individual letters (out of the context of a sentence), acronyms, indexes and footnotes were explicitly excluded from the samples. Headings, metalanguage, glossaries, contents of tables and word lists were included.

Defining the frequency of occurrence of the *corpus* items

The following thresholds were used as criteria for deciding whether a word had low, medium or high frequency: Low frequency – a probability lower than, or equal to, 0.008%; Medium frequency – greater than 0.008% and less than 0.02%; High frequency – equal to or greater than 0.02%. Table 3 shows the classification of the data by school year.

This classification generated, for each school year, a list of words organised in an Excel spreadsheet by increasing levels of frequency of occurrence. In each of these levels lexical items are presented in alphabetical order and each has the following information available: absolute frequency, number of letters and frequency per million tokens. The latter value is introduced in this new presentation of the *corpus* (Appendix 3) for the convenience of the reader (calculated as $1,000,000/tokens * (\text{relative frequency}) = 10,000/tokens * (\text{relative frequency as a \%})$). The relative frequency as a % is already shown in tables, one for each academic level, as can be seen in Appendix 4, which also show absolute frequency (also presented in Appendix 3) and additional descriptive data for each word.

As an example of the information in Appendix 4, the description of the data for the 1st school year shows that 1786 words occurred once, 734 twice, 452 three times and so on (See the 1st tab of the spreadsheet in Appendix 3). The relative frequency of the words that occur once is 0.00294. There are 1786 different words (*vocabulary*) that occur only once. To reiterate, the value for *vocabulary* refers to the number of different words (*types*) for each level of frequency. The quantity of words refers to the total number of words (*tokens*) and represents the value for *vocabulary* multiplied by its frequency of occurrence. As such, a word that occurs 8 times, for example, is counted as 8 words (*tokens*) and 1 vocabulary (*type*). It follows that vocabulary / word is a measure of the extent, or richness, of vocabulary. This measure is the total number for vocabulary divided by the total number of words.

Table 3.

The quantity of vocabulary (n) and the spread of the absolute frequency (AF) and relative frequency (RF) of the words belonging to the vocabulary of children from the 1st to the 5th school year banded as low, medium and high frequency

Year	Low frequency	Medium frequency	High frequency
1 st	n = 2957	n = 654	n = 947
	AF: 1–3	AF: 4–6	AF: >6
	RF: 0.00294–0.00881	RF: 0.01175–0.01762	RF: 0.02056–5.66309
2 nd	n = 10136	n = 818	n = 684
	AF: 1–17	AF: 18–41	AF: >41
	RF: 0.0048–0.00813	RF: 0.00860–0.01960	RF: 0.02007–3.93899
3 rd	n = 16696	n = 894	n = 645
	AF: 1–30	AF: 31–72	AF: >72
	RF: 0.00028–0.00826	RF: 0.00853–0.01982	RF: 0.02009–3.87102
4 th	n = 22871	n = 911	n = 650
	AF: 1–45	AF: 46–109	AF: >109
	RF: 0.00018–0.00816	RF: 0.00834–0.01976	RF: 0.01995–3.75970
5 th	n = 27093	n = 940	n = 622
	AF: 1–49	AF: 50–123	AF: >123
	RF: 0.00016–0.00795	RF: 0.00812–0.01996	RF: 0.02013–3.86490

Validity indices of the classification of the frequency of occurrence of items in Pinheiro's Word count (1996)

A recent study, Justi et al. (2013), set out to evaluate the generality and currentness of Pinheiro's *corpus* (1996) in a sample of 10% of the pages of five textbooks most commonly used by 2nd year Elementary school pupils of 95 municipal schools in the city of Maceió, Alagoas. In this sample the authors obtained a total of 9,672 *tokens* and 2,718 *types* (in the current *corpus* the number of tokens and types for the same school year are 209,216 and 11,640 respectively).

By means of Pearson's correlation analysis Justi et al. (2013) investigated whether there was a correlation between the absolute frequency of each *type* common to both Pinheiro (1996) and the sample obtained in Maceió. The strong correlation encountered ($r = 0.955$, $N = 1852$, $p < .001$) provided another validity indice for the current *corpus*, this despite the above cited research having a small sample size and being limited to one school year.

In fact, studies based on the original database, carried out in diverse parts of Brazil, at different times and with samples comprising different age groups, socio-economic levels and reading ability (ex., Godoy, 2005; Justi & Justi, 2009; Salles & Parente, 2002; Salles & Parente, 2007; Stivanin & Scheuer, 2005), have not only reported an effect on the frequency of reading and/or writing of the children tested, but that this effect is consistent with the expectations of the literature which, considering that it is the result of research carried out over 20 years, provides validity to the word count and its currentness.

Not only because of this validity, but also due to of the lexical variations in different sociolinguistic and regional contexts (that do not only apply to Brazilian Portuguese), this is a good place for a recommendation regarding the use of data contained in the current corpus in research conducted in different parts of Brazil, including in the State of Minas Gerais. To ensure the reliability of the results being obtained it is necessary that the researcher be satisfied that the words chosen for his/her research are in fact of high, medium or low frequency in his/her region. Therefore a pilot study should present the selected words to participants (a sample independent of the final study, but coming from the same population) along with other words (which can even be pseudowords) and ask them to assign a level of familiarity to each on a *likert* scale.

The relevance of the corpus and Pinheiro's Word count (1996)

Lists of words derived from a wide variety of sources, and containing in the region of two million words, are extremely useful to lexicographers, linguists, psycholinguists and educators for the better and more accurate analysis of language, for the development of research in the field of literacy learning, in the devising of literacy tests and for the development of evaluation and rehabilitation programmes for dyslexics and children with specific literacy problems.

Regarding psycholinguistics, the vocabulary of the current corpus could, depending on the aim of the user, be submitted to further analysis, beyond that presented here. It could, for example, be classified in terms of orthographic regularity, syllable structure and level of abstraction (concrete versus abstract words). With regard to orthographic regularity, Pinheiro (2007) classified the low frequency vocabulary common to children in the 2nd to 5th school years in terms of its grapheme-phoneme and phoneme-grapheme relationship in regular words, words governed by contextual and irregular words.

Bibliography

Carroll, J. B. (1972). Elementary English, Vol. 49, No. 7, pp. 1070-1074. National Council of Teachers of English. <http://www.jstor.org/stable/41387874> (accessed: 18/01/2015).

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American heritage word frequency book*. New York: American Heritage and Boston: Houghton Mifflin.

Godoy, D. M. A. (2005). *Aprendizagem inicial da leitura e da escrita no português do Brasil: influência da consciência fonológica e do método de alfabetização*. Tese de Doutorado, não publicada – Universidade Federal de Santa Catarina, Brasil.

Justi, C., & Justi, F. (2009). Os efeitos de lexicalidade, frequência e regularidade na leitura de crianças falantes do português brasileiro. *Psicologia: Reflexão e Crítica*, 22(2), 163-172.

Justi, F.R. dos R.; Justi, C. N. G., Almeida, M. F., Câmara (2013). A generalidade da contagem de frequência de ocorrência de palavras de Pinheiro (1996). In A. Roazzi, J. Salles F. R. dos R. Justi (Eds.), *A aprendizagem da leitura e da escrita: contribuições de pesquisa* (p. 53-66). São Paulo: Editora Vetor.

Micro-ocp (1988). *Oxford University computing service*. Oxford: Oxford University Press.

Pinheiro, A. M. V. (1991). Um levantamento sobre as publicações usadas nas redes pública e particular nas faixas pré-escolar e e escolar em Belo Horizonte. Monografia não publicada.

Pinheiro, A. M. V. (1996a). Word Frequency Count in Written Brazilian Portuguese. In S. Contento (Eds.). *Psycholinguistics as Multidisciplinarily Connected Science*. Cesena: Società Editrice Il Ponte Vecchio, 2, 47–52.

Pinheiro, A. M. V. (1996b). *Contagem de frequência de ocorrência de palavras expostas a crianças de 1ª à 4ª série do Ensino Fundamental*. São Paulo: Associação Brasileira de Dislexia.

Pinheiro, A. M. V. (2007). Anexo 2. In I. Sim-Sim & F. L. Vianna, *Para avaliação do desempenho da leitura* (pp. 119-130). Lisboa, Portugal: Gabinete de Estatística Ministério da Educação.

Salles, J. F., & Parente, M. A. M. P. (2002). Processos cognitivos na leitura de palavras em crianças: Relações com compreensão e tempo de leitura. *Psicologia: Reflexão e Crítica*, 15(2), 321-331.

Salles, J. F., & Parente, M. A. M. (2007). Avaliação da leitura e escrita de palavras em crianças de 2ª série: Abordagem neuropsicológica cognitiva. *Psicologia: Reflexão e Crítica*, 20(2), 220-228.

Sinclair, J. (1991). *Corpus concordance collocation*. Oxford: Oxford University Press.

Stivanin, L & Scheuer, C. I. Claudia Inês Scheuer Tempo de latência e exatidão para leitura e nomeação em crianças escolares: estudo piloto, *Educação e Pesquisa*, São Paulo, v. 31, n. 3, p. 425-436.