

## **Frequência de ocorrências de palavras em livros escolares expostos a crianças brasileiras nos anos iniciais do ensino fundamental**

Ângela Maria Vieira Pinheiro\*

Resumo: Apresenta-se aqui o *corpus* que se tornou conhecido na literatura brasileira como “A Contagem de Pinheiro (1996)”. Esse banco de itens foi desenvolvido com o objetivo de identificar a frequência de ocorrência de palavras no material *escrito* exposto às crianças do Estado de Minas Gerais, que cursavam os anos iniciais do ensino fundamental nas redes pública e particular, no período de 1990-1994. O exame de um amplo corpo de palavras (cerca de dois milhões), derivado de uma grande variedade de fontes, permitiu a identificação de um vocabulário (léxico) composto de palavras classificadas nas categorias de alta, média e baixa frequência de ocorrência, que é representativo do vocabulário ao qual as crianças são expostas durante períodos específicos do processo de aprendizagem da leitura e da escrita. Lexicógrafos, linguistas, psicolinguistas, educadores irão encontrar neste trabalho dados importantes que poderão ser utilizados para diferentes propósitos de pesquisa. No contexto do estudo da aprendizagem e desenvolvimento da leitura no Brasil, o presente *corpus* de palavras tem sido, desde que foi publicado, a principal fonte de referência para a pesquisa na área.

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.

\* Professora Titular. Departamento de Psicologia, Faculdade de Filosofia e Ciências Humanas (FAFICH), Universidade Federal de Minas Gerais, Brasil.

Contato: [pinheiroamva@gmail.com](mailto:pinheiroamva@gmail.com)

## **Introdução**

Em 1996, Pinheiro disponibilizou à comunidade científica brasileira um *corpus* sobre o vocabulário escrito exposto a crianças. Dada a importância desse trabalho, que desde a sua publicação tem sido referência para a pesquisa na área de desenvolvimento da leitura e da escrita, está sendo apresentado aqui na Plataforma CHILDES, acompanhado de informações e materiais atualizados relevantes para a sua compreensão e utilização.

## **Breve histórico**

O referido *corpus* foi elaborado por meio de dois estudos. O primeiro consistiu da seleção das publicações mais frequentemente usadas nos currículos da escola pré-primária e das séries iniciais (de 1ª. à 4ª série)<sup>1</sup> do Ensino Fundamental (EF) das redes pública (escolas municipais e estaduais) e particular da cidade de Belo Horizonte (Pinheiro, 1991). Esse estudo gerou o *input* para o segundo estudo, que, por sua vez, constitui-se da formação de um *corpus* de palavras, com subsequente contagem da frequência de ocorrência de cada uma delas (Pinheiro, 1996a, 1996b).

## **Estudo 1 – Seleção do input do *corpus***

O trabalho constou dos seguintes passos: 1) levantamento do número de escolas do 1º ao 5º ano do EF das redes de ensino de Belo Horizonte, seguido do sorteio estratificado de 10% delas em cada regional de ensino de cada rede; 2) elaboração do instrumento de investigação – um questionário que indagava sobre os livros-textos, livros suplementares e qualquer material escrito, mais usados nas áreas de Português, Matemática, Estudos Sociais e Ciências; 3) envio desse instrumento às supervisoras educacionais dos estabelecimentos selecionados para preenchimento; e 4) seleção de todas as publicações que receberam três ou mais citações.

---

<sup>1</sup> A pré-escola é equivalente ao 1º ano e as séries iniciais ao 2º ao 5º ano do EF atual de 9 anos, vigente no Brasil a partir 2006. Para propósito de atualização da informação, esta nomenclatura vigente será doravante utilizada.

Esse procedimento resultou na obtenção de 124 livros (8,19%, dos 1514 títulos diferentes levantados), sendo a maioria adotada em todos os estados do Brasil pelo Ministério da Educação. A distribuição desse material, por ano escolar, é apresentada na Tabela 1.

Tabela 1.

*Frequência de citação das publicações nas áreas de estudos investigadas do 1º. ao 5º. ano do EF\**

Área de estudo	Ano escolar					Total
	1º	2º	3º	4º	5º	
Português	8	13	13	11	11	56
Matemática	4	8	7	8	6	33
Estudos Soc./Ciências	4	6	7	9	9	35
<b>Total</b>	16	27	27	28	26	124

\*Publicações com suas respectivas referências são apresentadas nos Anexos 1 e 2.

## Estudo 2 – Formação do *corpus* de palavras

Os textos extraídos das publicações selecionadas no Estudo 1, após serem digitados e editados (com o consentimento formal de suas respectivas editoras), foram processados pelo Oxford Concordance Program (OCP), o qual registrou 1774164 palavras (quantidade de palavras contidas no *corpus* ou *tokens*) das 124 publicações, originando um vocabulário de 40509 palavras (número de palavras diferentes ou *types*). A distribuição desses dados, por ano escolar, é apresentada na Tabela 2.

O critério para a entrada de um item no *corpus* foi a sua ocorrência no texto em termos de *graphic word type*, uma vez que o OCP podia somente reconhecer letras individuais, sequências de letras e caracteres. Assim, essa contagem, como a de Carrol, Davis e Richman (1971), trata *acabar*, *acaba*, *acabada*, *acabado*, *acabam* (e as demais formas derivadas do morfema (radical) *acab-*), por exemplo, como diferentes entradas e para cada uma delas oferece a sua ocorrência. Como explica Carrol (1972), “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). Tomando o exemplo acima, pode-se descobrir a frequência combinada de todas as palavras com radical *acab-*.

No entanto, fugindo ao critério descrito, foram consideradas como unidades lexicais individuais as ocorrências constituídas por uma sequência de *graphic word types* tais como as formas onomatopaicas (ex., *pam ram pam pam*), interjeições (ex., *ufa ufa*) e palavras dando noção de melodia ou qualquer som inventado (ex., *la la la la*). Os itens pertencentes a essa categoria foram digitados sem o espaço entre seus componentes (ex., *lalalala*) e subsequentemente, no processo editorial das listas geradas pelo OCP, os espaços foram colocados de volta em seus respectivos lugares.

O hífen foi mantido apenas nas palavras compostas e nas formas verbais em que há supressão da consoante final do verbo (ex., *cortá-la*, *conhecê-la*). A retirada do hífen nesse tipo de construção gramatical daria origem a palavras não existentes na língua: “*cortá*” e “*conhecê*”. Seguindo esse raciocínio, nas formas gramaticais tais como *viu-se*, *comeu-se* e *deu-me*, nas quais cada um dos dois itens unidos pelo hífen são palavras por si só, os hífen foram removidos.

Também, como na contagem de Carrol, Davis e Richman (1971), todas as palavras com igual grafia foram tratadas como a mesma palavra, independentemente dos usos e diferentes significados que possam ter em contextos diferentes. Consequentemente, itens do tipo *banco*, *canto* e *mato* (que podem ter a função de verbo ou substantivo comum) foram considerados como unidades não diferenciadas. O acesso à complexidade gramatical e semântica dessas palavras só pode ser alcançado por meio de consulta ao contexto nos quais ocorrem (informação não disponível no presente *corpus*).

Finalmente, numerais (0, 1, 2, etc.), letras individuais (fora do contexto de sentenças), acrônimos, o conteúdo de índices e de notas de rodapé foram explicitamente excluídos das amostras de texto. Foram mantidos os títulos, metalinguagem, glossários e os conteúdos de tabelas e listas de palavras.

Tabela 2.

*Distribuição do número de palavras registradas (tokens) e de vocabulário por ano escolar (types)*

Ano escolar	<i>Tokens</i>	Types
1º	34045	4573
2º	209216	11640
3º	363313	18297
4º	551506	24521
5º	616084	28742
TOTAL	1774164	40509*

\*Valor excluiu as palavras duplicadas nos diferentes anos escolares.

#### *Definição da frequência de ocorrência dos itens corpus*

Utilizou-se, como critério para decisão do valor limite para a consideração de uma palavra com tendo baixa, média ou alta frequência, a probabilidade de ocorrência inferior ou igual a 0.008%, entre 0.008% e 0.02%, igual ou superior a 0.02%, respectivamente. A Tabela 3 apresenta essa classificação para os dados do 1º ano até o 5º ano.

Essa classificação gerou para cada nível escolar uma lista de palavras organizada em uma planilha do excel (bloqueada para edição) em nível crescente de frequência de ocorrência: palavras de baixa frequência, seguidas das de frequência média e essas por palavras de alta frequência. Em cada um desses níveis, os itens lexicais são apresentados em ordem alfabética e para cada uma deles os seguintes dados são disponibilizados: frequência absoluta, número de letras e frequência por milhão, sendo este último valor, introduzido nesta nova apresentação do *corpus* (Anexo 3) por conveniência para o leitor (calculado como  $1000000/tokens * (\text{frequência relativa}) = 10000/tokens * (\text{frequência relativa em } \%)$ ). Já a frequência relativa (em porcentagem) é mostrada em tabelas, uma para cada nível

acadêmico, conforme pode ser visto no Anexo 4, que também exibe a frequência absoluta (igualmente apresentada no Anexo 3) e dados descritivos adicionais para cada palavra.

Tabela 3.

*Quantidade de vocabulário (n), e amplitude da frequência absoluta (FA) e relativa (FR) das palavras pertencentes ao vocabulário de crianças do 1º ano até o 5º ano dentro das faixas baixa, média e alta frequência*

<b>Ano</b>	<b>Baixa frequência</b>	<b>Média frequência</b>	<b>Alta frequência</b>
1º	n = 2957	n = 654	n = 947
	FA: 1–3	FA: 4–6	FA: >6
	FR: 0.00294–0.00881	FR: 0.01175–0.01762	FR: 0.02056–5.66309
2º	n = 10136	n = 818	n = 684
	FA: 1–17	FA: 18–41	FA: >41
	FR: 0.0048–0.00813	FR: 0.00860–0.01960	FR: 0.02007–3.93899
3º	n = 16696	n = 894	n = 645
	FA: 1–30	FA: 31–72	FA: >72
	FR: 0.00028–0.00826	FR: 0.00853–0.01982	FR: 0.02009–3.87102
4º	n = 22871	n = 911	n = 650
	FA: 1–45	FA: 46–109	FA: >109
	FR: 0.00018–0.00816	FR: 0.00834–0.01976	FR: 0.01995–3.75970
5º	n = 27093	n = 940	n = 622
	FA: 1–49	FA: 50–123	FA: >123
	FR: 0.00016–0.00795	FR: 0.00812–0.01996	FR: 0.02013–3.86490

Como ilustração das informações oferecidas no Anexo 4, a descrição dos dados do 1º ano indica que 1786 palavras ocorreram uma vez, 734 duas vezes, 452 três vezes e assim por diante (1ª aba da planilha do Anexo 3). A frequência relativa das palavras que ocorrem uma vez é 0.00294. Existem 1786 palavras diferentes (vocabulário) que ocorreram apenas 1 vez. Reiterando, a quantidade de “vocabulário” refere-se então ao número de palavras diferentes (*types*), em cada nível de frequência. Já a quantidade de palavras refere-se ao número total de palavras (*tokens*) e representa o número de vocabulário multiplicado por

sua frequência de ocorrência. Assim, uma palavra que ocorre 8 vezes, por exemplo, é contada como 8 palavras (*tokens*) e 1 vocabulário (*type*). A razão vocabulário/palavra é uma medida da extensão ou riqueza do vocabulário. Essa medida é a divisão do número total do vocabulário pelo número total de palavras.

### **Índices de validade da classificação de frequência de ocorrência dos itens da Contagem de Pinheiro (1996)**

Em um estudo recente, Justi et al. (2013) propuseram avaliar a generalidade e a atualidade do *corpus* de Pinheiro (1996) em uma amostra de 10% das páginas de cinco livros didáticos mais usados por alunos do 2º ano do EF de 95 escolas municipais da cidade de Maceió, Alagoas. Dessa amostra, os autores obtiveram um total de 9672 *tokens* e 2718 *types* (no presente *corpus*, o número de *tokens* e *types* para o mesmo ano escolar é 209216 e 11640, respectivamente).

Por meio de análises de correlação de *Pearson*, Justi et al. (2013) investigaram se a frequência absoluta para cada *type* comum entre Pinheiro (1996) e a da amostra de obtida em Maceió se correlacionam. A forte correlação encontrada,  $r = .955$ ,  $N = 1852$ ,  $p < .001$ , proveu mais um índice de validade para o presente *corpus*, a despeito da pequena amostra de textos da pesquisa supracitada e de se limitar a apenas a um nível escolar.

De fato, estudos com base no banco de dados original, realizados em diversas partes do Brasil, em diferentes épocas e com amostras compreendendo várias faixas etárias, nível sócio-econômico e habilidade de leitura (ex., Godoy, 2005; Justi & Justi, 2009; Salles & Parente, 2002; Salles & Parente, 2007; Stivanin & Scheuer, 2005), não só têm reportado um efeito de frequência na leitura e/ou escrita das crianças testadas, mas que esse efeito está de acordo com as expectativas da literatura, o que oferece validade para a referida contagem de palavras e também para a sua atualidade, considerando que ela é resultante de pesquisa conduzida há mais de 20 anos.

Por essa razão, e também devido às variações do léxico em diferentes contextos sociolinguísticos e regionais, o que se aplica, não só ao português brasileiro, cabem aqui algumas recomendações a respeito da utilização dos dados contidos no presente *corpus* em pesquisas realizadas em diferentes partes do Brasil (incluindo o Estado de Minas Gerais). Para assegurar a confiabilidade dos resultados a serem obtidos, é necessário que o

pesquisador se certifique de que as palavras escolhidas para a sua pesquisa sejam, de fato, de alta, média ou de baixa frequência, em sua região. Portanto, em um estudo piloto, deve apresentá-las aos participantes (uma amostra independente do estudo final, mas oriunda da mesma população), juntamente com outras palavras (que podem ser inclusive pseudopalavras) e lhes solicitar que assinalem, em uma escala *likert*, o nível de familiaridade de cada uma.

### **Relevância do Corpus e da Contagem de Pinheiro (1996)**

Listas de palavras derivadas de uma grande variedade de fontes, contendo cerca de dois milhões de palavras, são de grande utilidade para lexicógrafos, linguistas, psicolinguistas e educadores, para uma melhor e mais adequada análise da língua; para o desenvolvimento de pesquisas na área de aprendizagem da leitura/escrita; elaboração de testes de leitura/escrita e finalmente, para o desenvolvimento de programas de avaliação e reeducação para disléxicos e crianças com problemas específicos na leitura/escrita.

No que se refere à psicolinguística, o vocabulário do presente *corpus* pode, dependendo do propósito do usuário, ser submetido a outras análises, além das já apresentadas. Pode, por exemplo, ser classificado em termos de regularidade ortográfica, estrutura silábica e nível de abstração (palavras concretas *versus* abstratas). No que se refere à regularidade ortográfica, Pinheiro (2007) classificou o vocabulário de baixa frequência comum às crianças do 2º ao 5º ano em termos de sua relação grafema-fonema e fonema-grafema em palavras regulares, regidas por regras contextuais e irregulares.

### **Referências**

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 Wtitten 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.